Original Paper

Engagement, Feasibility, and Perceived Alliance With a Large Language Model–Enabled Mental Health Voice and Chatbot: Exploratory Longitudinal Observational Study

Lea Maria Schäfer[1], BA, BSc, MSc; Tabea Krause[1], MSc; Stephan Köhler[1,2], Prof Dr

[1]Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[2]Department of Psychiatry, Psychotherapy and Psychosomatics, Alexianer St. Joseph-Krankenhaus Berlin-Weißensee, Berlin, Germany

Corresponding Author:

Lea Maria Schäfer, BA, BSc, MSc
Department of Psychiatry and Psychotherapy
Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin
Charitéplatz 1
Berlin 10117
Germany
Phone: +49 17645760761
Email: lea.schaefer@charite.de

Abstract

Background: Conversational AI systems are increasingly used for web-based mental health support, yet evidence on sustained engagement, therapeutic alliance, and longitudinal outcomes remains limited. Baseline data from 527 self-referred users of the large language model–enabled chatbot Clare indicated high psychological distress and early development of a subjective working alliance.

Objective: This exploratory study examined the feasibility and descriptive outcome patterns associated with the use of the multimodal conversational agent Clare over 4 and 8 weeks in a real-world setting, focusing on engagement, perceived working alliance, and mental health measures.

Methods: We conducted a naturalistic, single-arm, repeated-measures observational study of self-referred, English-speaking adults recruited online via social media and search advertisements. Participants interacted with Clare over 8 weeks. Two analytic samples were defined by engagement and assessment completion: a 4-week sample (active engagement plus completed 4-week assessment) and an 8-week sample (additionally requiring continued engagement and completed 8-week assessment). Outcomes were assessed at baseline, 4, and 8 weeks: engagement metrics (calls, messages, and call duration), alliance measures (Working Alliance Inventory–Short Revised and Mobile Agnew Relationship Measure), and mental health measures (Patient Health Questionnaire–4, Patient Health Questionnaire–Distress, Mini–Social Phobia Inventory, UCLA Loneliness Scale, and Satisfaction With Life Scale). Changes were evaluated using 2-tailed paired-samples *t* tests with effect sizes; given substantial attrition and reliance on complete-case analyses, linear mixed-effects models were run as sensitivity analyses, and findings warrant cautious interpretation.

Results: Of 527 baseline participants, 369 enrolled and 348 actively engaged with Clare; 53 out of 348 (15.2%) met the 4-week criteria and 21 out of 348 (6.0%) met the 8-week criteria. Engagement peaked in week 1 (calls: mean 1.77, SD 1.52; messages: mean 10.02, SD 16.68) and declined thereafter. Among T1-T2 complete pairs (n=52), mean Patient Health Questionnaire–Distress scores decreased by 1.13 (95% CI 0.25 to 2.02) points, Patient Health Questionnaire–4 scores decreased by 0.69 (95% CI –0.01 to 1.40) points, Mini–Social Phobia Inventory scores decreased by 1.73 (95% CI 0.88 to 2.58) points, Satisfaction With Life Scale scores increased by 6.23 (95% CI 4.70 to 7.76) points, and UCLA scores increased by 0.48 (95% CI 0.00 to 0.96) points. Among participants with Working Alliance Inventory–Short Revised data (n=53), total scores increased by 0.20 points from T1b to T2 (95% CI –0.01 to 0.42); among the 8-week retained subgroup (n=21), Mobile Agnew Relationship Measure scores increased by 0.22 points from T2 to T3 (95% CI 0.03 to 0.42).

Conclusions: Sustained engagement was limited, and outcome estimates were obtained from small, highly selected follow-up samples. The observed changes cannot be attributed to Clare because the study had no control group and substantial, nonrandom attrition.

JMIR Form Res 2026;10:e89708; doi: 10.2196/89708

Keywords: conversational agent; large language models; mental health bots; working alliance; digital mental health; voice-based interaction; longitudinal observational study; feasibility; artificial intelligence; therapeutic alliance

Introduction

Background

Depression, anxiety, and loneliness remain highly prevalent worldwide, yet many individuals do not receive adequate psychological support due to stigma, cost, and limited access [1]. Even among those who receive care, fewer than half benefit substantially from current interventions, underscoring the need for scalable mental health solutions [2-5].

Conversational AI (CAI) tools have been studied as scalable approaches to mental health support and self-management of depression, anxiety, and well-being [6-8]. Advances in speech recognition, natural language processing (NLP), and machine learning allow conversational agents (CAs) to assess clinical information and adapt interventions [9]. However, limited empathy, inconsistent alliance, and high dropout remain concerns [10-12].

Recent CAI systems are increasingly powered by large language models (LLMs), which can be adapted or fine-tuned for specific applications, including mental health support. Multimodal CAI systems that integrate both text-based and voice-based interactions may help address these limitations by fostering stronger engagement and therapeutic alliance [8,13,14]. Emerging work on "synthetic relationships" similarly suggests that continuous, adaptive interactions may support therapeutic processes [15], a critical factor for effective psychological treatment [13,16].

Evidence from naturalistic, voice-based use remains sparse [10], and little is known about the persistence of the perceived alliance. Prior studies have reported associations between baseline symptom severity or loneliness and subsequent outcomes in AI-supported interventions [17-20].

Relational characteristics, including attachment anxiety, epistemic trust, fear of intimacy, and personality traits, have also been associated with willingness to adopt CAI-based support [21-23]. These findings raise questions about how selection into CAI use relates to sustained engagement and observed outcomes [17].

A prior baseline study of 527 self-referred users of the Clare mental health chatbot addressed part of this question by showing that individuals often engage with CAs under conditions of considerable emotional and social vulnerability. That baseline report focused on cross-sectional data, describing participants' motives, expectations, and bonding with Clare (Working Alliance Inventory–Short Revised [WAI-SR]), as well as a broad set of psychosocial and demographic variables. Most users reported elevated loneliness (86%), anxiety (69%), and depression (59%), and many cited avoiding embarrassment or concerns about appearance in face-to-face therapy as reasons for choosing a chatbot. Users also reported a strong perceived working alliance with Clare (WAI-SR; mean 3.76, SD 0.72) [14].

The baseline analysis did not examine longitudinal interaction patterns, perceived alliance, or psychological measures. The present study uses the same cohort to address these questions [14].

Aim

This exploratory feasibility study described engagement, perceived alliance, and psychological measures over 4 to 8 weeks of real-world Clare use. It characterized modality preferences and explored associations between baseline psychological profiles, subsequent engagement, alliance, and symptom scores. The analyses were descriptive and hypothesis-generating.

Methods

The Intervention: Clare Voice- and Chatbot

Clare is an AI-based CA developed by clare&me GmbH, now owned by HealthyCloud GmbH. The evaluated version was deployed during the study's data collection period (November 2023 to April 2024) and combined a symptom-to-intervention matching algorithm, proprietary NLP-based safeguarding mechanisms for mental health use, and a self-hosted, fine-tuned open-source Llama LLM.

The LLM was trained on a synthetic data set iteratively developed and reviewed by psychologists and psychotherapists, in collaboration with licensed clinicians, to translate evidence-based cognitive behavioral therapy (CBT) elements into supportive conversations. With user consent, conversational data and derived recommendations may be stored to enable personalized support via a selective memory function [24]. Clare is intended to complement, not replace, professional mental health care and does not provide diagnosis or emergency services.

Minor software updates during the study did not affect the underlying LLM, therapeutic content, or safeguarding mechanisms; further technical details cannot be disclosed for proprietary reasons.

Structure, Content, and Delivery of Clare

Clare provides psychological support through voice-based and text-based conversations grounded in CBT, mindfulness, and self-compassion. CBT is a structured, evidence-based psychotherapeutic approach that targets cognitive and behavioral processes through techniques such as Socratic questioning and cognitive restructuring [14,24].

Interactions followed a semistructured conversational flow—an emotional check-in, identification of current

concerns, and delivery of targeted interventions (eg, breathing exercises, cognitive restructuring, psychoeducation, or sleep hygiene), and users could engage on demand via phone or text, switching flexibly between modalities (Figure 1).

Figure 1. User interface of a typical Clare session: (A) incoming voice call from Clare, (B) text-based interaction with Clare, and (C) transition from texting to a voice call following a user request. During the study, participants could initiate a call either by requesting Clare to call them or by scheduling a call via a study-specific website provided during onboarding.



Intervention content was selected through a rule-based system developed by clinical psychologists, mapping user input (eg, expressions of stress or fatigue) to tailored coping strategies and adapting responses based on interaction history. Voice interactions tended to be more guided and linear, while text interactions allowed greater flexibility and user-directed pacing; sessions typically lasted 5 to 45 minutes depending on engagement and intervention type.

Onboarding & Exposure to Clare

All participants completed an onboarding call covering Clare's purpose, scope, limitations, and instructions for use, and were informed that it was an AI-based CA—not a human therapist or substitute for psychotherapy—and not intended for emergency or crisis care. Participants could engage with Clare on demand via phone or text, switch flexibly between modalities, and use it for as long as they wished (Figure 1). Registration and onboarding initiation constituted enrollment, while exposure was defined as any interaction beyond the onboarding call; inclusion in the analytic samples required completing the onboarding call and relevant assessments, plus at least 1 interaction between assessment time points. Interaction metrics captured the frequency and duration of calls and text-based interactions.

Safety System of Clare

The system incorporates rule-based safety mechanisms, content filtering, and automated risk assessment to identify safety-relevant content, including suicidality, self-harm, harm to others, and elevated clinical symptom burden. When risk indicators are detected, the system may initiate clarifying questions to assess severity before providing an empathic response, communicating the limitations of AI support, and directing users to appropriate professional services and crisis resources in their country. These safety protocols were developed and reviewed in collaboration with practicing psychologists to support the ethical management of risk-related disclosures.

Conversations were not monitored by clinicians in real time; instead, risk detection and response were managed through the automated safety system. Conversation data could be reviewed retrospectively for quality assurance, safety evaluation, and system improvement in accordance with user consent. Further details on Clare's design and safety framework are reported in Schäfer et al [14].

Study Design and Settings

This naturalistic, single-arm, repeated-measures observational study examined real-world use of Clare among self-referred users, drawing on data from the Clare baseline study. Participants independently signed up and engaged with the intervention outside a controlled research setting. The primary aim was to explore engagement patterns, feasibility, and preliminary associations rather than evaluating causal efficacy. Participants interacted with the multimodal CA, which supported both voice and text communication, over an 8-week period.

Data were collected via SoSciSurvey between November 20, 2023, and April 21, 2024. Assessments occurred at sign-up prior to intervention use (T1), 3 to 5 days after initial engagement (T1b), week 4 (T2), and week 8 (T3); participants could complete questionnaires after the scheduled date, so T2 and T3 refer to the scheduled rather than actual completion time points. Because registration did not necessarily indicate active use, retention and longitudinal analyses were restricted to participants meeting engagement criteria established in the baseline study.

Participants

Participants were recruited between November 2023 and April 2024 through English-language advertisements on Facebook, Instagram, LinkedIn, and Google Ads, using

language targeting based on users' browser or account settings. No demographic targeting was applied. The sampling frame comprised adults interested in using a digital mental health intervention who had access to a mobile or internet-enabled device.

Participants received detailed information on the study and completed the baseline survey (T1) via SoSciSurvey after providing informed consent. Eligible participants were aged 18 to 65 years, fluent in English, and had access to a mobile internet-enabled device. Detailed baseline characteristics of this cohort have been reported previously [14].

Analytic Sample

To ensure meaningful exposure to the intervention, analyses were restricted to participants who progressed beyond enrollment and actively engaged with Clare. Enrollment was defined as registration and initiation of the onboarding process, whereas active engagement required completion of the first postengagement assessment (T1b) and at least one voice- or chat-based interaction between assessment periods. Analyses further included only participants who met the predefined engagement criteria and did not report suicidality at onboarding or during follow-up, as described previously [14]. No formal sample size calculation was conducted; sample size was determined by feasibility within the recruitment period.

Two analytic samples were defined. The 4-week sample (T2; n=53) included participants who completed T1b and T2. The 8-week sample (T3; n=21) included participants who completed all assessments from T1 to T3. Based on the engagement criteria, participants were required to complete at least two interactions for inclusion in the 4-week sample and at least three interactions for inclusion in the 8-week sample, in addition to the onboarding call. The analytic samples therefore represent a subset of enrolled and actively engaged participants and may not be representative of all registered users. Further details are provided in Multimedia Appendix 1.

Safety Monitoring Procedures

In addition to the automated safety mechanisms described, anonymized conversation transcripts were reviewed twice weekly by 2 members of the research team for quality assurance and safety monitoring. Reviews focused on identifying potential failures of risk detection, inappropriate or clinically inaccurate responses, and other safety-relevant issues. Suicidality detection and sensitive-topic dialogue management were implemented as dedicated, rule-governed components.

Safety monitoring supported ongoing system refinement; however, formal evaluation of safety performance metrics (eg, false-positive and false-negative rates) was not performed as part of this observational study.

The technical design was rooted in responsible and ethical guidelines on the development of AI in mental health [25,26] and psychological practice [27], determining when the system should de-escalate or redirect users to human assistance.

Ethical Considerations

The study followed an observational design without randomization or a control group. The study protocol was approved by the Ethics Committee of Campus Charité Mitte, Berlin, Germany (EA1/109/22), and the procedures complied with applicable institutional and national ethical standards. Study data were collected and managed independently by the Charité research team; HealthyCloud GmbH did not have access to the deidentified research data set. All participants provided informed consent prior to participation. Participant data were deidentified before analysis and handled in accordance with applicable data protection regulations. Participants received €15 (€1=US $1.07 as of April 19, 2024) as compensation upon completion of the survey and were allowed to continue using Clare free of charge after the study.

Measures

While newer approaches such as ecological momentary assessment (EMA) [28,29] offer real-time symptom assessment and could be integrated with AI-based interventions, we used traditional self-report measures to ensure consistency with prior digital mental health research.

Outcomes included psychological distress (Patient Health Questionnaire–4 [PHQ-4] and Patient Health Questionnaire–Distress [PHQ-D], including the stress module), perceived working alliance (WAI-SR), mobile therapeutic alliance (Mobile Agnew Relationship Measure [mARM]), loneliness (UCLA Loneliness Scale [UCLA]), life satisfaction (Satisfaction With Life Scale [SWLS]), and AI engagement metrics (call frequency, duration, and modality). All measures except the mARM were assessed at T1, T1b, T2 (week 4), and T3 (week 8), with T1b serving as the analytic baseline for outcomes requiring prior interaction; the mARM was assessed only at T2 and T3, reflecting the assumption that therapeutic relationship development requires sustained engagement.

Situational context variables were collected only at T3. We report depression, anxiety, distress, loneliness, and social anxiety, as these constructs are most commonly examined in AI mental health studies and allow for comparability with existing evidence [6,19,30,31]. Except for the mARM, all measures are described in detail in Schäfer et al [14].

Behavioral Engagement Constructs

Data were collected automatically through the web-based interface of the Clare platform during user interactions. These data included engagement metrics such as the number of calls, call duration, and modality of interaction (voice or text), as well as timestamps of usage. Engagement data (onboarding call initiation, texting, and calling) were captured via automated logs of user interactions with the system, without manual data entry by researchers. A more detailed description of the captured engagement data is available in our baseline study report [14].

Situational Context of Use and User Feedback

At the 8-week assessment (T3), participants reported the circumstances in which they used Clare using predefined response options (eg, at home, on the go, during a crisis, as part of a routine, or at specific times of day). Participants were also asked whether they would recommend Clare for self-help purposes. In addition, an optional open-ended question invited participants to provide feedback regarding their experience with the intervention and study procedures. Open-ended responses were reviewed descriptively and summarized narratively; no formal qualitative analysis was conducted.

Measures of Alliance With the AI

WAI-SR

Perceived working alliance was assessed with the 12-item WAI-SR, adapted by replacing "therapist" with "Clare" as in the baseline study [14,32 The scale yields scores from 1 to 5 across bond, goal, and task dimensions, with values ≥3.45 indicating high alliance.

mARM

The mARM [33] is a 25-item adaptation of the Agnew Relationship Measure for use in mobile health (mHealth) interventions. It assesses 5 components of perceived therapeutic alliance: bond, partnership, confidence, openness, and client initiative. Items are rated on a 7-point Likert scale ranging from 1 ("strongly disagree") to 7 ("strongly agree").

Consistent with prior research [34-36], scores ≥5 indicate a strong alliance, scores around 4 reflect a moderate relationship, and scores <4 may suggest a weak alliance. The Agnew Relationship Measure and mARM have been widely used in digital mental health research. Following established adaptation procedures, references to "the app" in the mARM were replaced with "Clare" to reflect the context of the present study; further details are provided in Table S1 in Multimedia Appendix 2.

Mental Health Constructs

PHQ-4

The PHQ-4, combining the Patient Health Questionnaire–2 and 2-item Generalized Anxiety Disorder scale, assesses depression and anxiety on a 4-point scale, with subscale scores ≥3 indicating a positive screen and total scores (0-12) reflecting none (0-2), mild (3-5), moderate (6-8), or severe (9-12) psychological distress [37].

PHQ-D Module (PHQ-D Stress)

The PHQ distress module (part of the PHQ-D) assesses psychosocial stress across 10 items rated 0-2, with total scores (0-20) indicating minimal (0-4), mild (5-9), moderate (10-14), or severe (15-20) stress [37].

Three-Item Social Phobia Inventory

The Mini–Social Phobia Inventory (Mini-SPIN) is a 3-item short form of the Social Phobia Inventory that screens for social anxiety by assessing fear, avoidance, and related physiological symptoms. Items are rated from 0 ("not at all") to 4 ("extremely"), and scores of ≥6 typically indicate elevated social anxiety [38].

Three-Item UCLA Loneliness Scale

The 3-item UCLA Loneliness Scale assesses loneliness and social isolation, with items rated from 1 to 4 and averaged to yield a total score, where higher values indicate greater loneliness [39].

SWLS

The 5-item SWLS is a brief measure of global life satisfaction. Items are rated on a 7-point scale (1="strongly disagree" to 7="strongly agree"), producing total scores from 5 to 35, with a score of 20 suggesting a moderate level of life satisfaction and higher scores reflecting greater life satisfaction [40].

Data Analysis

Analyses examined engagement patterns and changes in mental health and bonding measures across T1, T1b, T2, and T3. Two analytic subsamples were used: a 4-week follow-up sample (T2; n=53) and an 8-week follow-up sample (T3; n=21), both consisting of participants who met the predefined engagement and assessment completion criteria. Categorical engagement data were summarized using frequencies; open-ended responses were summarized narratively. To examine interaction modality, participants were categorized based on their primary mode of engagement with the intervention. Descriptive statistics are presented in Table 1. "Voice-only" users engaged exclusively via phone calls, "text-only" users interacted exclusively through written messages, and "hybrid" users used a combination of both modalities. As the majority of participants engaged with the intervention using a hybrid modality, modality-specific descriptive analyses primarily focused on this group to maximize the number of observations available for interpretation.

Table 1. Descriptive statistics for mental health and well-being measures at baseline (T1), by subsequent follow-up completion.

Measure	4-week follow-up sample[a] (n=53), mean (SD)	8-week follow-up sample[b] (n=21), mean (SD)
PHQ-D[c]	10.47[d] (3.92)	8.76 (3.37)
PHQ-4[e]	6.25[f] (3.49)	5.05 (2.96)
Mini-SPIN[g]	7.74[h] (3.29)	7.05 (2.97)
UCLA[i]	6.21[j] (1.98)	5.81 (2.20)

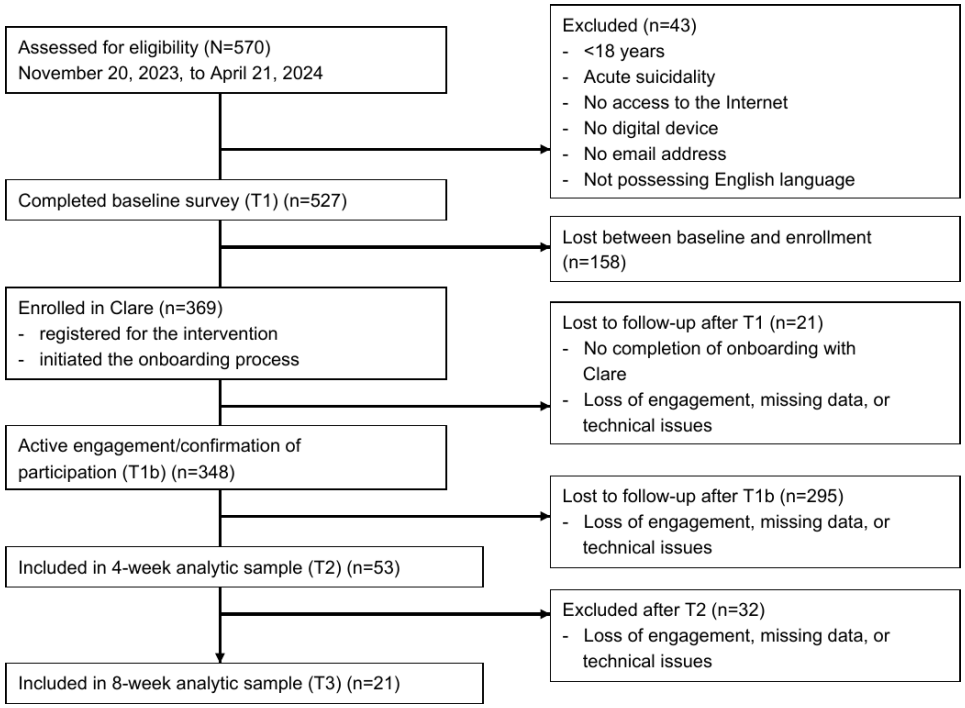
Measure	4-week follow-up sample[a] (n=53), mean (SD)	8-week follow-up sample[b] (n=21), mean (SD)
SWLS[k]	19.81[l] (6.65)	19.52 (5.90)

[a]Values represent T1 scores for participants who subsequently completed the T2 (n=53) assessment.
[b]Values represent T1 scores for participants who subsequently completed the T3 (n=21) assessment.
[c]PHQ-D: Patient Health Questionnaire–Distress.
[d]Scores 10-14 reflect moderate to severe stress 15-20.
[e]PHQ-4: Patient Health Questionnaire–4.
[f]Scores >3 reflect clinical symptoms of anxiety or depression.
[g]Mini-SPIN: Mini–Social Phobia Inventory.
[h]Scores ≥6 are classified as at risk of a generalized social anxiety disorder.
[i]UCLA: UCLA Loneliness Scale.
[j]Scores ≥6 is classified as "lonely" (UCLA).
[k]SWLS: Satisfaction With Life Scale.
[l]Scores ≥10 classify as (extremely) dissatisfied; scores ≥26 as (extremely) satisfied with one's life.

Quantitative analyses were conducted in SPSS version 28. Paired samples 2-tailed *t* tests compared WAI-SR scores between T1b and T2 and between T1b and T3, and mARM scores between T2 and T3. For all other mental health outcomes, comparisons were made between T1 and T2 and between T1 and T3. Mean differences and 95% CIs were reported as estimation-based summaries, with Cohen *d* values and *P* values reported only in tables.

Retention was defined based on continued engagement with Clare rather than registration alone, as sign-up did not necessarily indicate active use of the intervention. Participant retention was tracked across 3 key time points: T1b (first engagement-based assessment), T2 (weeks 1-4), and T3 (weeks 1-8). T1b reflects an early postengagement assessment rather than the baseline assessment. Participant progression from baseline assessment (n=527) to enrollment (n=369), active engagement (n=348), and analytic inclusion at T2 (n=53) and T3 (n=21) is illustrated in Figure 2.

Figure 2. STROBE (Strengthening the Reporting of Observational Studies in Epidemiology)–style participant flow diagram showing recruitment, eligibility screening, engagement with Clare, retention, and inclusion in longitudinal analyses. Participant flow from baseline survey (n=527) through Clare onboarding and follow-up assessments. Boxes indicate baseline completers, enrollees, actively engaged users, and participants retained at 4-week (T2; n=53) and 8-week (T3; n=21) follow-up, which formed the analytic samples for longitudinal analyses.



Given the exploratory nature of the study, no correction for multiple comparisons was applied. This approach was chosen to maintain sensitivity for detecting potentially meaningful associations and to reduce the risk of Type II errors, particularly given the limited sample size at follow-up.

Missing Data, Attrition, and Sensitivity Analyses

Given the high attrition rates, primary analyses were based on complete cases. To assess the missing data mechanism, Little's test for missing completely at random (MCAR) was performed. Because the data were not assumed to be

MCAR, sensitivity analyses were conducted using linear mixed-effects models with restricted maximum likelihood. Models included time (T1, T2, and T3) as a fixed effect, a random intercept for participant, and a first-order autoregressive covariance structure to account for repeated observations within individuals. All available observations from participants contributing data at one or more assessment points were included; complete data across all waves were not required. The number of participants and observations contributing to each model varied by outcome due to attrition and outcome-specific missingness. No data imputation was performed.

To characterize attrition, baseline demographic, clinical, and alliance variables were compared between completers and noncompleters using the 2-tailed Welch *t* tests and descriptive statistics. The Welch *t* tests were selected because the assumption of homogeneity of variance was violated. Effect sizes were calculated using Cohen *d* (Multimedia Appendix 3).

Results

Retention and Dropout Analysis

Participant flow is shown in Figure 2. At T1b, 348 participants remained actively engaged. Of these, 53 (15.2%) met the T2 analytic criteria and 21 (6.0%) met the T3 criteria, corresponding to 94.0% attrition from T1b to T3.

Between-group estimates and 95% CIs for baseline differences between completers and noncompleters are provided in Multimedia Appendix 3.

Safety Outcomes

No instances of self-harm or suicidality were identified by the automated risk assessment system during the study period, and no participants were excluded or withdrawn because of safety-related concerns. Transcript reviews did not identify any safety-critical system responses requiring participant follow-up or clinical intervention.

Sample Characteristics of Analytic Samples

The T2 (n=53) and T3 (n=21) analytic samples were demographically similar. The mean ages of participants in the T2 (n=53) and T3 (n=21) analytic samples were 33.19 (SD 9.74; range 18-55) years and 34.86 (SD 10.65; range 19-55) years, respectively, and most were based in the United Kingdom or the United States. Sociodemographic characteristics across baseline (T1), early postengagement (T1b), 4-week (T2), and 8-week (T3) samples are listed in Table 2.

Table 2. Sociodemographic characteristics across baseline (T1), early postengagement (T1b), 4-week (T2), and 8-week (T3) samples.

Variable	T1 (n=527)	T1b (n=348)	T2 (n=53)	T3 (n=21)
Age (years), mean (SD; range)	36.2 (9.39; 18-64)	36.94 (9.31; 18-63)	33.19 (9.74; 18-55)	34.86 (10.65; 19-55)
Gender, n (%)				
Women	277 (52.56)	176 (50.57)	39 (73.58)	15 (71.43)
Men	245 (46.49)	168 (48.28)	11 (20.75)	5 (23.81)
Other/prefer not to say	5 (0.95)	4 (1.15)	3 (5.66)	1 (4.76)
Country, n (%)				
United Kingdom	207 (39.28)	161 (46.26)	27 (50.94)	11 (52.38)
United States	129 (24.48)	100 (28.74)	13 (24.53)	4 (19.05)
Germany	158 (29.98)	77 (22.13)	0 (0%)	3 (14.29)
Other	33 (6.26)	10 (2.87)	13 (24.53)	3 (14.29)
Ethnicity, n (%)				
White	397 (75.33)	262 (75.29)	32 (60.38)	13 (61.90)
African American	53 (10.06)	42 (12.07)	3 (5.66)	0 (0)
Asian	30 (5.69)	19 (5.46)	12 (22.64)	5 (23.81)
Hispanic	20 (3.80)	9 (2.59)	1 (1.89)	0 (0%)
Mixed	20 (3.80)	12 (3.45)	2 (3.77)	1 (4.76)
Other	5 (0.95)	2 (0.57)	1 (1.89)	0 (0)
I chose not to say	2 (0.38)	2 (0.57)	2 (3.77)	2 (9.52)
Employment status, n (%)				
Employed	443 (84.06)	299 (85.92)	40 (75.47)	14 (66.67)
Unemployed	33 (6.26)	22 (6.32)	8 (15.09)	5 (23.81)
Retired	24 (4.55)	19 (5.46)	1 (1.89)	0 (0)
Other	27 (5.12)	8 (2.30)	4 (7.55)	2 (9.52)
Relationship status, n (%)				
Single	204 (38.71)	126 (36.21)	28 (52.83)	12 (57.14)

Variable	T1 (n=527)	T1b (n=348)	T2 (n=53)	T3 (n=21)
Married	255 (48.39)	182 (52.30)	16 (30.19)	5 (23.81)
Other	68 (12.90)	40 (11.49)	9 (16.98)	4 (19.05)
Highest education, n (%)				
Middle school	18 (3.42)	46 (13.22)	3 (5.66)	2 (9.52)
Apprenticeship	28 (5.31)	11 (3.16)	2 (3.77)	1 (4.76)
High school/A-level	295 (55.98)	115 (33.05)	15 (28.30)	5 (23.81)
University degree	165 (31.31)	154 (44.25)	28 (52.83)	11 (52.38)
Other	21 (3.98)	22 (6.32)	5 (9.43)	2 (9.52)
Prior digital mental health tool use, n (%)[a]	8 (1.52)	4 (1.15)	4 (7.55)	4 (19.05)

[a]Reflects the number and percentage of participants answering "yes" to this question.

Mental Health at Baseline

Table 2 displays baseline (T1) scores for depression, anxiety, distress, social anxiety, loneliness, and life satisfaction for the 4-week (n=53) and 8-week (n=21) analytic subsamples. Values fall within comparable ranges across groups and reflect moderate distress and social anxiety alongside relatively low life satisfaction.

Interaction Modality Preference (Voice, Text, or Hybrid)

Hybrid use was the most frequently recorded modality among retained participants, whereas voice-only use was rare (Table 3).

Table 3. Participant numbers by time point and mode of interaction[a].

Time frame	Value, n	Text-only use (no additional calls), n (%)	Voice-only use, n (%)	Hybrid, n (%)
T1b–T2 (4-week sample)	53	17 (32.1)	0 (0.0)	36 (67.9)
T1b–T3 (8-week sample)	21	5 (23.8)	2 (9.5)	14 (66.7)

[a]Participants were included if they completed the relevant assessments and engaged with the intervention at least once between assessment points (in addition to the onboarding call). Percentages are calculated within each time frame. Interaction modes refer to usage beyond onboarding: text-only use (no additional calls), voice-only use (phone-based interactions only), and hybrid (use of both modalities).

Situational Context of Use and User Satisfaction Signals

After 8 weeks, participants (n=21) indicated the contexts in which they used the CAI (multiple responses allowed). Most reported using it at home (18/21, 85.7%), whereas only 2 (9.5%) participants used it "on the go," and none reported use during an acute crisis. Regarding timing, 9 (42.9%) participants used the CAI as part of a routine, whereas 11 (52.4%) used it flexibly. Of 21 participants, 8 (38.1%) used it during the day, and 5 (23.8%) at night. Most participants stated they would recommend Clare for self-help (18/21, 85.7%).

Written feedback was provided by 10 (47.6%) participants, with most comments expressing satisfaction with the CAI or the questionnaire (eg, "Everything was good," "Very clear and no problems encountered," and "I am satisfied with Clare"); some pointed to minor ambiguities (eg, distinguishing in-person from online therapy). The feedback received was generally positive; however, the limited number of comments and the absence of a formal qualitative analysis preclude firm conclusions regarding usability.

Descriptive Use Patterns: Calling and Texting Behavior

Engagement metrics across weeks 1 to 4 are presented in Table 4. Engagement was highest in week 1 and declined consistently across all indicators over the study period.

Table 4. Descriptive statistics for engagement metrics across weeks 1 to 4 (n=53)[a].

Week	Calls, mean (SD)	Messages, mean (SD)	Call length (minutes), mean (SD)
1	1.77 (1.52)	10.02 (16.68)	3.35 (4.37)
2	0.68 (0.96)	6.17 (10.02)	2.01 (3.23)
3	0.49 (0.64)	4.74 (7.97)	1.25 (2.55)
4	0.40 (0.57)	2.72 (4.65)	1.45 (2.95)
Weeks 1-4 overall	3.34 (2.53)	23.65 (32.82)	3.33 (3.49)

[a]Overall call and message values represent cumulative counts per participant. Overall call length represents participants' average call durations, calculated by weighting weekly average durations by weekly call counts. Values are mean (SD) across participants.

Working Alliance and Mental Health Measurements

Changes in working alliance and mental health outcomes were examined for the 4-week (n=53) and 8-week (n=21) analytic samples; sample sizes varied slightly across outcomes due to missing data. Longitudinal trajectories are presented in Table 5 and Table 6. Additional analyses are provided in Multimedia Appendix 4 and Multimedia Appendix 3.

Table 5. Paired-sample comparisons of primary outcomes between baseline (T1) and follow-up (T2; n=53 and n=52 for mental health outcomes)[a].

Measurement[a]	T1, mean (SD)	T2, mean (SD)	*t* test[b] (*df*)	*P* value	Cohen *d*[c]	95% CI[d]
PHQ-D[e]	10.42 (3.94)	9.29 (3.75)	2.56 (51)	.01	0.36	0.07 to 0.63
PHQ-4[f]	6.23 (3.52)	5.54 (3.20)	1.97 (51)	.054	0.27	–0.01 to 0.55
Mini-SPIN[g]	7.79 (3.30)	6.06 (3.86)	4.10 (51)	<.001	0.57	0.27 to 0.86
SWLS[h]	19.67 (6.64)	25.90 (7.38)	–8.16 (51)	<.001	–1.13	–1.48 to –0.78
UCLA[i]	6.19 (2.00)	6.67 (2.01)	–2.03 (51)	.048	–0.28	–0.56 to 0.00
WAI-SR[j] total[k]	3.20 (0.91)	3.40 (1.02)	–1.88 (52)	.07	–0.26	–0.53 to 0.02
Goal	3.15 (1.05)	3.32 (1.16)	–1.32 (52)	.19	–0.18	–0.45 to 0.09
Task	3.04 (0.92)	3.27 (1.14)	–1.82 (52)	.08	–0.25	–0.52 to 0.02
Bond	3.39 (1.00)	3.60 (0.98)	–1.80 (52)	.08	–0.25	–0.52 to 0.03

[a]Degrees of freedom vary due to missing data and reflect the number of paired observations. *t* statistics and Cohen *d* were computed as T1 (or T1b) minus T2; negative values indicate an increase in scores over time and positive values indicate a decrease.
[b]2-tailed *t* test.
[c]Cohen *d* for paired samples.
[d]CI for effect size.
[e]PHQ-D: Patient Health Questionnaire–Distress.
[f]PHQ-4: Patient Health Questionnaire–4.
[g]Mini-SPIN: Mini–Social Phobia Inventory.
[h]SWLS: Satisfaction With Life Scale.
[i]UCLA: UCLA Loneliness Scale.
[j]WAI-SR: Working Alliance Inventory–Short Revised.
[k]Goal, Task, Bond, and WAI-SR total comparisons are based on the early postengagement assessment (T1b) and T2.

Table 6. Longitudinal changes in symptom, well-being, and alliance measures over 8 weeks of Clare engagement (n=21).

Measurement[a]	T1, mean (SD)	T2, mean (SD)	T3, mean (SD)
PHQ-4[b]	5.05 (2.96)	4.95 (2.56)	3.86 (2.83)
Anxiety subscale	2.62 (1.72)	2.47 (1.60)	2.00 (1.70)
Depression subscale	2.43 (1.66)	2.47 (1.43)	1.86 (1.53)
PHQ-D[c]	8.76 (3.37)	8.38 (2.87)	8.14 (3.15)
Mini-SPIN[d]	7.05 (2.97)	5.38 (3.06)	5.48 (2.94)
SWLS[e]	19.52 (5.90)	24.90 (7.16)	21.67 (5.59)
UCLA[f]	5.81 (2.20)	6.43 (2.18)	5.38 (1.43)
WAI-SR[g] total[h]	2.91 (0.88)	3.21 (1.17)	3.34 (1.03)
Goal	2.65 (1.04)	3.05 (1.33)	3.21 (1.16)
Task	2.88 (0.93)	3.07 (1.25)	3.20 (1.13)
Bond	3.20 (0.97)	3.51 (1.07)	3.61 (0.95)
mARM[i]	—[j]	4.73 (0.78)	4.96 (0.68)

[a]Possible score ranges were PHQ-D=0-27; PHQ-4 total=0-12 (subscales 0-6 each); Mini-SPIN=0-12; UCLA=3-9; SWLS=5-35; mARM=0-7. Higher scores indicate greater symptom severity (PHQ-D, PHQ-4, Mini-SPIN, and UCLA), greater life satisfaction (SWLS), and stronger perceived alliance (WAI-SR and mARM).
[b]PHQ-4: Patient Health Questionnaire–4.
[c]PHQ-D: Patient Health Questionnaire–Distress.
[d]Mini-SPIN: Mini–Social Phobia Inventory.
[e]SWLS: Satisfaction With Life Scale.
[f]UCLA: UCLA Loneliness Scale.
[g]WAI-SR: Working Alliance Inventory–Short Revised.
[h]Goal, Task, Bond, and WAI-SR total comparisons are based on the early postengagement assessment (T1b) and T2.
[i]mARM: Mobile Agnew Relationship Measure.
[j]Not applicable; the mARM was assessed only at T2 and T3, reflecting the assumption that therapeutic relationship development requires sustained engagement.

Longitudinal Changes in Adapted Working Alliance and Bonding Measures

From T1b to T2 (n=53), mean WAI-SR total scores increased by 0.20 (95% CI –0.01 to 0.42) points. Bond scores at T2 exceeded the ≥3.45 threshold proposed by Jasper et al [41] for human therapeutic contexts. Among the retained 8-week subgroup (n=21), mARM scores increased by 0.22 points from T2 to T3 (95% CI 0.03 to 0.42; Table 5 and Table 6).

Longitudinal Changes in Distress, Depressive, and Anxiety Symptoms

Among T1–T2 complete pairs (n=52), mean PHQ-D scores decreased by 1.13 (95% CI 0.25 to 2.02) points, PHQ-4 scores decreased by 0.69 (95% CI –0.01 to 1.40) points, and Mini-SPIN scores decreased by 1.73 (95% CI 0.88 to 2.58) points (Table 5).

Longitudinal Changes in Life Satisfaction (SWLS)

Among T1-T2 complete pairs (n=52), mean SWLS scores increased by 6.23 (95% CI 4.70 to 7.76) points (Table 5).

Longitudinal Changes in Loneliness (UCLA)

Among T1–T2 complete pairs (n=52), mean UCLA scores increased by 0.48 (95% CI 0.00 to 0.96) points. Among the 8-week subgroup, the descriptive means were 5.81 at T1, 6.43 at T2, and 5.38 at T3 (Table 6).

Changes in Mental Health, Well-Being, and Bonding Across 8 Weeks

Table 6 presents descriptive trajectories for the 21 participants retained through T3. These estimates describe this selected subgroup and do not estimate an intervention effect.

Relationships Between Alliance, Distress, and User Engagement

Associations between early communication frequency, alliance, and mental health measures at T2 were explored in the 4-week sample (n=53). Full estimates are provided in Multimedia Appendix 5.

Associations of Alliance With Distress, Loneliness, and Engagement

Correlations between WAI-SR and depression, anxiety, loneliness, or social anxiety had absolute values of 0.20 or less. Message-volume correlations with WAI-SR ranged from ϱ of –0.36 (approximate 95% CI –0.57 to –0.10) to ϱ of –0.28 (approximate 95% CI –0.50 to –0.01); call-frequency and call-duration estimates were closer to zero (Multimedia Appendix 5).

Sensitivity Analysis Findings

Little's MCAR assessment indicated that missingness was not completely at random. Sensitivity analyses using all available observations produced changes in the same direction as the primary complete-case analyses for WAI-SR, PHQ-D, Mini-SPIN, and SWLS, whereas the mARM estimate indicated little change. Thus, the overall pattern broadly converged with the primary analyses, although attrition-related bias remains possible; detailed model outputs are provided in Multimedia Appendix 6.

Discussion

Principal Findings

Sustained engagement was limited: 53 out of 348 (15.2%) actively engaged users met the 4-week criteria and 21 (6.0%) met the 8-week criteria. Hybrid voice-and-text use was common among retained participants, whereas voice-only use was rare. In the complete-case samples, mean alliance and several symptom or well-being scores changed over time, but wide or boundary-crossing CIs for several estimates, the absence of a control group, and the highly selected follow-up samples limit interpretation.

Higher baseline distress and loneliness were associated with lower subsequent engagement. Alliance estimates were only weakly associated with communication volume or symptom measures, indicating that perceived bonding and usage intensity should not be treated as interchangeable indicators. These findings describe associations among retained users and do not establish effectiveness, causal mechanisms, or the clinical importance of the observed changes.

Comparison With Prior Work

Engagement and Attrition in Digital Mental Health Interventions

The sharp early decline in engagement is consistent with attrition reported for unguided conversational-agent and digital mental health interventions [12,42-45]. Across CA studies, reported dropout has averaged approximately 22% and ranged from 0% to 83%, depending in part on intervention duration and study design [7,42]. Retention in the present study was at the low end of that literature, illustrating the difficulty of sustaining engagement with an autonomous, unguided agent delivered at scale.

Sustaining engagement without increasing cognitive or practical burden remains a broader challenge for digital health interventions [46]. Technical friction and repetitive or stylistically predictable LLM output may contribute to disengagement [35,47-51]. Clare used rule-based pathways and contextual adaptation intended to reduce repetition; however, perceived repetition and technical barriers were not systematically assessed and therefore cannot be linked to dropout in this study.

Interaction modality may also shape continued use. Text and voice differ in pacing, privacy, effort, and conversational richness [8,52,53]. Most retained participants combined text and calls and used Clare mainly at home, but the small voice-only group precluded modality-specific inference. The

observed preference for hybrid use, therefore, describes this selected sample rather than demonstrating that one modality is more acceptable or effective.

Engagement is also defined inconsistently across digital mental health research. Existing work highlights the lack of standardized metrics, limited evidence linking usage intensity to outcomes, and insufficient user involvement in defining meaningful engagement [54,55]. Frequency of use, sustained participation, perceived benefit, and relational experience may represent different constructs. Future studies should also consider possible overreliance, emotional dependency, and problematic use rather than assuming that greater use is necessarily preferable.

Baseline Distress and Subsequent Engagement

Higher baseline distress was associated with lower subsequent engagement, consistent with evidence that users with greater symptom burden may discontinue digital interventions more often [42]. However, meta-analytic findings regarding differences between general, at-risk, and clinical samples have been mixed [12,50]. The present association should therefore be treated as exploratory and may reflect symptom-related motivation, interface burden, unmeasured life circumstances, or other selection factors [56,57].

Participants mainly used Clare during routine activities rather than acute crises, which is compatible with a role in ongoing self-management rather than crisis support [58]. The findings support further study of tailored onboarding, optional human support, and consent-based proactive contact for users with greater symptom burden [10]. They do not show that these approaches would improve retention or outcomes.

User Engagement and Working Alliance

Engagement declined after initial use, consistent with novelty-related patterns in digital health and human–machine interaction [45,59,60]. Perceived alliance scores changed only modestly among retained users. Therapeutic alliance traditionally refers to agreement on goals and tasks together with an affective bond and mutual trust [16,61,62]. Applying this construct to a nonsentient agent therefore requires conceptual caution.

Although WAI-SR and mARM have been used or adapted in digital contexts [33,63-65], it remains uncertain whether they capture equivalent relational processes with AI. Ratings may reflect users' perceptions, expectations, anthropomorphism, or satisfaction rather than a reciprocal therapeutic partnership. The combination of declining engagement and relatively stable alliance ratings is compatible with literature describing human–AI relationships as psychologically meaningful but qualitatively distinct from human therapeutic relationships [15,33].

Communication volume showed little association with perceived alliance. Observational work with Woebot users has similarly identified heterogeneous profiles in which lower use did not uniformly correspond to lower alliance [66,67]. These patterns suggest that usage frequency and relational perception should be measured separately. They also underscore the need for AI-specific validation of alliance measures and for research on how perceived alliance relates to retention and clinical outcomes.

Symptom Changes in Unguided or Autonomous CAI Intervention

Retained participants entered the study with moderate distress and elevated social anxiety and loneliness, a profile described in other digital help-seeking samples [18,19]. Changes in symptom and well-being scores were estimated only among participants with follow-up data. They may reflect selective retention, regression to the mean, assessment-related response shifts, concurrent life events, or natural symptom variation rather than an intervention-related change.

Prior studies have reported that baseline distress, fear of intimacy, epistemic trust, and other relational characteristics are associated with conversational-AI uptake and outcomes [17-21]. The present data do not establish which users benefit from autonomous agents or whether stepped-care models combining automated and human support would improve outcomes. Distinguishing characteristics associated with initial uptake from those associated with sustained use or benefit remains an important research priority.

Loneliness increased at T2 and decreased in the much smaller T3 subgroup. This trajectory may reflect measurement variability or selective attrition and should not be interpreted as a temporal effect. Given broader concerns that digital relationships could either support connection or contribute to social withdrawal for some users [68,69], future studies should assess social functioning and offline relationships alongside loneliness.

Design Implications for AI-Supported Mental Health Interventions

Future design studies should test whether symptom-sensitive onboarding, adaptive support, and light-touch human guidance improve retention. Modality preference and early engagement signals may help identify when users need a different interaction style or additional support, but the present exploratory associations are insufficient for defining adaptation rules. Such strategies should be evaluated prospectively rather than inferred from post hoc subgroup patterns [10,17,21,70].

Any proactive or adaptive system requires explicit consent, transparent operation, user control, safety escalation, and clinical oversight [25,26]. Design research should include users with different levels of symptom burden and should examine both desired outcomes and possible harms, including dependency, overreliance, privacy concerns, and displacement of human support.

Limitations and Future Directions

The main limitation was substantial, nonrandom attrition, which produced small and highly selected follow-up samples. Participants who discontinued had higher psychological

distress at baseline [14]. Complete-case estimates may therefore reflect survivorship bias, regression to the mean, or natural variation and may overrepresent less distressed or more motivated users. Mixed-effects sensitivity analyses incorporated all available observations but cannot remove bias when dropout is related to unobserved outcomes.

The uncontrolled design precludes causal attribution and does not separate changes associated with Clare use from expectancy, repeated assessment, concurrent care, life events, or the passage of time. Generalizability is further limited by a predominantly female, self-referred sample, financial compensation for survey completion, and the small number of participants retained beyond 4 weeks. Future studies should use adequately powered controlled designs, including appropriate active chatbot controls where feasible.

Reliance on self-report measures and 4-week assessment intervals may introduce recall, response, or social-desirability bias. EMA could provide more temporally sensitive information [28,29]. Alliance measures were developed outside AI relationships and may not capture their dynamic or nonreciprocal nature. Validated AI-specific measures of trust, anthropomorphism, expectations, and perceived relational quality are needed.

Interpretation of engagement was constrained by minimal voice-only use and by the requirement for an onboarding call, which may have discouraged users preferring purely text-based support [71]. Message and call frequency do not capture conversation quality, reasons for use, or the user's intended level of engagement. The study also lacked systematic qualitative analysis of disengagement and user experience.

No formal adverse-event ascertainment or expert evaluation of LLM response accuracy, appropriateness, empathy, or safety was conducted. Although Clare incorporated automated safety mechanisms and referral procedures, the study cannot determine how consistently these features performed in naturalistic use. Future evaluations should combine structured safety monitoring, expert review, qualitative work, and assessment against clinical guidance.

Future research should prioritize retention strategies developed with users, more diverse samples, longer follow-up, and prespecified estimation-based analyses. Studies should compare autonomous and human-supported delivery, text, voice, and hybrid formats, and therapeutic and nontherapeutic controls. Research should distinguish factors associated with initial uptake, sustained engagement, perceived alliance, and clinical change rather than treating them as a single pathway [18,63,72,73].

Conclusions

In this uncontrolled naturalistic study, engagement declined substantially and only small, selected subgroups completed the 4-week and 8-week assessments. Hybrid use was common among those retained, and perceived alliance and several mental health scores changed among completers, but the estimates cannot be attributed to Clare. The divergence between usage and perceived alliance highlights the need for clearer, AI-appropriate definitions and measures of relational engagement. Controlled, adequately powered studies with stronger retention, systematic safety evaluation, and longer follow-up are needed.

Acknowledgments

The authors would like to thank all participants who engaged with the Clare chatbot and contributed to this study. The support of the clare&me GMBH and HealthyCloud GmbH team involved in the development and maintenance of the Clare system is also gratefully acknowledged. LMS worked with clare&me GmbH throughout the project and was affiliated with clare&me GmbH during the study. clare&me GmbH is now part of HealthyCloud GmbH, where LMS continues to work part-time. The authors declare the use of generative AI (GenAI) in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing, summarizing text, adapting and adjusting emotional tone, and reformatting. The GenAI tool used was GPT-5.2 (OpenAI). Responsibility for the final manuscript lies entirely with the authors; GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

HealthyCloud GmbH contributed to the publication fees for this article.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to their sensitive nature (mental health data), applicable data protection regulations (eg, General Data Protection Regulation), and their proprietary status involving a commercial partner. The data include potentially identifiable information related to user interactions with the Clare chat- and voice-bot, including conversation transcripts and associated mental health assessments. Because conversation transcripts contain highly sensitive free-text mental health disclosures that cannot be reliably anonymized without compromising participant privacy, they cannot be shared. Deidentified questionnaire data and other nontranscript study data may be made available from the corresponding author on reasonable request, subject to approval by the company, the relevant ethics committee, and in accordance with institutional and legal requirements.

Authors' Contributions

LMS contributed to conceptualization, methodology, investigation, data curation, and formal analysis, and led the writing of the manuscript, including drafting the original version and integrating coauthor feedback. TK contributed to data curation, formal analysis, and critical review of the analyses. SK (senior author) contributed to supervision and critically reviewed and edited the manuscript. All authors reviewed and approved the final manuscript.

Conflicts of Interest

LMS was employed by clare&me GmbH, which developed the Clare voice and chatbot evaluated in this study, during the study period. LMS currently works part-time for HealthyCloud GmbH, which now owns Clare. LMS was also involved in the development of the intervention. All other authors declare no conflicts of interest.

Multimedia Appendix 1

Demographic characteristics of the T2 and T3 analytic samples.
[DOCX File (Microsoft Word File), 6739 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Items of the Working Alliance Inventory–Short Revised (WAI-SR) questionnaire, in which the word "the app" was changed to "Clare" in adaptation for this study.
[DOCX File (Microsoft Word File), 7164 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Dropout analysis.
[DOCX File (Microsoft Word File), 6742 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Additional data of the posttreatment sample (n=21).
[DOCX File (Microsoft Word File), 6738 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Correlations of therapeutic alliance scores (Mobile Agnew Relationship Measure [mARM]) at T2 (4 weeks) and T3 (8 weeks) with mental health variables (loneliness, social anxiety, Patient Health Questionnaire–Distress [PHQ-D], and Patient Health Questionnaire–4 [PHQ-4]) as well as engagement with Clare (total numbers of calls, average call length, and total number of messages).
[DOCX File (Microsoft Word File), 7075 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Sensitivity analyses.
[DOCX File (Microsoft Word File), 6738 KB-Multimedia Appendix 6]

References

1. Waumans RC, Muntingh ADT, Draisma S, Huijbregts KM, van Balkom AJLM, Batelaan NM. Barriers and facilitators for treatment-seeking in adults with a depressive or anxiety disorder in a Western-European health care setting: a qualitative study. BMC Psychiatry. Mar 5, 2022;22(1):165. [doi: 10.1186/s12888-022-03806-5] [Medline: 35247997]
2. Cuijpers P, Cristea IA, Karyotaki E, Reijnders M, Huibers MJH. How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. World Psychiatry. Oct 2016;15(3):245-258. [doi: 10.1002/wps.20346] [Medline: 27717254]
3. Cuijpers P, Miguel C, Harrer M, et al. Cognitive behavior therapy vs. control conditions, other psychotherapies, pharmacotherapies and combined treatment for depression: a comprehensive meta-analysis including 409 trials with 52,702 patients. World Psychiatry. Feb 2023;22(1):105-115. [doi: 10.1002/wps.21069] [Medline: 36640411]
4. Schaeuffele C, Meine LE, Schulz A, et al. A systematic review and meta-analysis of transdiagnostic cognitive behavioural therapies for emotional disorders. Nat Hum Behav. Mar 2024;8(3):493-509. [doi: 10.1038/s41562-023-01787-3] [Medline: 38228727]
5. Simon GE, Moise N, Mohr DC. Management of depression in adults: a review. JAMA. Jul 9, 2024;332(2):141-152. [doi: 10.1001/jama.2024.5756] [Medline: 38856993]
6. Edge D, Watkins ER, Limond J, Mugadza J. The efficacy of self-guided internet and mobile-based interventions for preventing anxiety and depression - a systematic review and meta-analysis. Behav Res Ther. May 2023;164:104292. [doi: 10.1016/j.brat.2023.104292] [Medline: 37003138]

7. He Y, Yang L, Qian C, et al. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res. Apr 28, 2023;25:e43862. [doi: 10.2196/43862] [Medline: 37115595]
8. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med. Dec 19, 2023;6(1):236. [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]
9. Diaz-Asper C, Hauglid MK, Chandler C, Cohen AS, Foltz PW, Elvevåg B. A framework for language technologies in behavioral research and clinical applications: ethical challenges, implications, and solutions. Am Psychol. Jan 2024;79(1):79-91. [doi: 10.1037/amp0001195] [Medline: 38236217]
10. Deng Y, Liao L, Zheng Z, Yang GH, Chua TS. Towards human-centered proactive conversational agents. Presented at: SIGIR '24: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval; Jul 14-18, 2024; Washington, DC, USA. [doi: 10.1145/3626772.3657843]
11. Salah M, Abdelfattah F, Al Halbusi H. The good, the bad, and the GPT: reviewing the impact of generative artificial intelligence on psychology. Curr Opin Psychol. Oct 2024;59:101872. [doi: 10.1016/j.copsyc.2024.101872] [Medline: 39197407]
12. Torous J, Lipschitz J, Ng M, Firth J. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. J Affect Disord. Feb 15, 2020;263:413-419. [doi: 10.1016/j.jad.2019.11.167] [Medline: 31969272]
13. Dosovitsky G, Bunge EL. Bonding with bot: user feedback on a chatbot for social isolation. Front Digit Health. 2021;3:735053. [doi: 10.3389/fdgth.2021.735053] [Medline: 34713203]
14. Schäfer LM, Krause T, Köhler S. Exploring user characteristics, motives, and expectations and the therapeutic alliance in the mental health conversational AI Clare: a baseline study. Front Digit Health. 2025;7:1576135. [doi: 10.3389/fdgth.2025.1576135] [Medline: 40585405]
15. Starke C, Ventura A, Bersch C, et al. Risks and protective measures for synthetic relationships. Nat Hum Behav. Oct 2024;8(10):1834-1836. [doi: 10.1038/s41562-024-02005-4] [Medline: 39438683]
16. Baier AL, Kline AC, Feeny NC. Therapeutic alliance as a mediator of change: a systematic review and evaluation of research. Clin Psychol Rev. Dec 2020;82:101921. [doi: 10.1016/j.cpr.2020.101921] [Medline: 33069096]
17. Herbener AB, Damholdt MF. A theoretical framework of the processes of change in mental health interventions delivered by artificial therapists. J Psychol AI. Dec 31, 2025;1(1):2590495. [doi: 10.1080/29974100.2025.2590495]
18. Chiauzzi E, Williams A, Mariano TY, et al. Demographic and clinical characteristics associated with anxiety and depressive symptom outcomes in users of a digital mental health intervention incorporating a relational agent. BMC Psychiatry. Jan 30, 2024;24(1):79. [doi: 10.1186/s12888-024-05532-6] [Medline: 38291369]
19. Kim M, Lee S, Kim S, et al. Therapeutic potential of social chatbots in alleviating loneliness and social anxiety: quasi-experimental mixed methods study. J Med Internet Res. Jan 14, 2025;27(1):e65589. [doi: 10.2196/65589] [Medline: 39808786]
20. Sobowale K, Humphrey DK. Evaluating the quality of psychotherapy conversational agents: framework development and cross-sectional study. JMIR Form Res. Jul 2, 2025;9(1):e65605. [doi: 10.2196/65605] [Medline: 40600851]
21. Guglielmucci F, Di Basilio D. Predicting engagement with conversational agents in mental health therapy by examining the role of epistemic trust, personality, and fear of intimacy: cross-sectional web-based survey study. JMIR Hum Factors. Jul 30, 2025;12(1):e70698. [doi: 10.2196/70698] [Medline: 40737594]
22. Heng S, Zhang Z. Attachment anxiety and problematic use of conversational artificial intelligence: mediation of emotional attachment and moderation of anthropomorphic tendencies. Psychol Res Behav Manag. 2025;18:1775-1785. [doi: 10.2147/PRBM.S531805] [Medline: 40873934]
23. Wu X, Liew K, Dorahy MJ. Trust, anxious attachment, and conversational AI adoption intentions in digital counseling: a preliminary cross-sectional questionnaire study. JMIR AI. Apr 22, 2025;4:e68960. [doi: 10.2196/68960] [Medline: 40262137]
24. Schäfer LM. Generative KI in der Psychotherapie: Chancen, Risiken und ethische Überlegungen – am Beispiel des Bots Clare [Article in German]. Verhaltensther Psychosoz Prax. 2025;57:390-407. URL: https://www.dgvt-verlag.de/zeitschriften/vpp3-2025-s5-390407/ [Accessed 2026-10-09]
25. Leslie D. Understanding artificial intelligence ethics and safety: a guide for the responsible design and implementation of AI systems in the public sector. SSRN. Preprint posted online on Jun 10, 2019. [doi: 10.2139/ssrn.3403301]
26. Stade EC, Stirman SW, Ungar LH, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. Npj Mental Health Res. Feb 2, 2024;3(1):12. [doi: 10.1038/s44184-024-00056-z]
27. Artificial intelligence and the field of psychology. American Psychological Association (APA); Aug 2024. URL: https://www.apa.org/about/policy/statement-artificial-intelligence.pdf [Accessed 2026-10-09]

28. De la Barrera U, Arrigoni F, Monserrat C, Montoya-Castilla I, Gil-Gómez JA. Using ecological momentary assessment and machine learning techniques to predict depressive symptoms in emerging adults. Psychiatry Res. Feb 2024;332:115710. [doi: 10.1016/j.psychres.2023.115710] [Medline: 38194800]
29. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol. 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
30. Dosovitsky G, Pineda BS, Jacobson NC, Chang C, Escoredo M, Bunge EL. Artificial intelligence chatbot for depression: descriptive study of usage. JMIR Form Res. Nov 13, 2020;4(11):e17065. [doi: 10.2196/17065] [Medline: 33185563]
31. Martinengo L, Lum E, Car J. Evaluation of chatbot-delivered interventions for self-management of depression: content analysis. J Affect Disord. Dec 15, 2022;319:598-607. [doi: 10.1016/j.jad.2022.09.028] [Medline: 36150405]
32. Hatcher RL, Gillaspy JA. Development and validation of a revised short version of the Working Alliance Inventory. Psychother Res. Jan 2006;16(1):12-25. [doi: 10.1080/10503300500352500]
33. Berry K, Salter A, Morris R, James S, Bucci S. Assessing therapeutic alliance in the context of mHealth interventions for mental health problems: development of the mobile Agnew relationship measure (mARM) questionnaire. J Med Internet Res. Apr 19, 2018;20(4):e90. [doi: 10.2196/jmir.8252] [Medline: 29674307]
34. Clarke J, Proudfoot J, Whitton A, et al. Therapeutic alliance with a fully automated mobile phone and web-based intervention: secondary analysis of a randomized controlled trial. JMIR Ment Health. Feb 25, 2016;3(1):e10. [doi: 10.2196/mental.4656] [Medline: 26917096]
35. Kay-Lambkin F, Baker A, Lewin T, Carr V. Acceptability of a clinician-assisted computerized psychological intervention for comorbid mental health and substance use problems: treatment adherence data from a randomized controlled trial. J Med Internet Res. Jan 27, 2011;13(1):e11. [doi: 10.2196/jmir.1522] [Medline: 21273184]
36. Grodniewicz JP, Hohol M. Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. Front Psychiatry. 2023;14:1190084. [doi: 10.3389/fpsyt.2023.1190084] [Medline: 37324824]
37. Kroenke K, Spitzer RL, Williams JBW, Löwe B. An ultra-brief screening scale for anxiety and depression: the PHQ–4. Psychosomatics. 2009;50(6):613-621. [doi: 10.1176/appi.psy.50.6.613] [Medline: 19996233]
38. Connor KM, Davidson JR, Churchill LE, Sherwood A, Foa E, Weisler RH. Psychometric properties of the Social Phobia Inventory (SPIN). New self-rating scale. Br J Psychiatry. Apr 2000;176:379-386. [doi: 10.1192/bjp.176.4.379] [Medline: 10827888]
39. Hughes ME, Waite LJ, Hawkley LC, Cacioppo JT. A short scale for measuring loneliness in large surveys. Res Aging. Nov 2004;26(6):655-672. [doi: 10.1177/0164027504268574] [Medline: 18504506]
40. Diener E, Emmons RA, Larsen RJ, Griffin S. The Satisfaction with Life Scale. J Pers Assess. Feb 1985;49(1):71-75. [doi: 10.1207/s15327752jpa4901_13] [Medline: 16367493]
41. Jasper K, Weise C, Conrad I, Andersson G, Hiller W, Kleinstäuber M. The working alliance in a randomized controlled trial comparing Internet-based self-help and face-to-face cognitive behavior therapy for chronic tinnitus. Internet Interv. Apr 2014;1(2):49-57. [doi: 10.1016/j.invent.2014.04.002]
42. Jabir AI, Lin X, Martinengo L, Sharp G, Theng YL, Tudor Car L. Attrition in conversational agent–delivered mental health interventions: systematic review and meta-analysis. J Med Internet Res. Feb 27, 2024;26(1):e48168. [doi: 10.2196/48168] [Medline: 38412023]
43. Lee YC, Yamashita N, Huang Y. Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional. Proc ACM Hum Comput Interact. May 29, 2020;4:1-27. [doi: 10.1145/3392836]
44. Nwosu A, Boardman S, Husain MM, Doraiswamy PM. Digital therapeutics for mental health: is attrition the Achilles heel? Front Psychiatry. 2022;13:900615. [doi: 10.3389/fpsyt.2022.900615] [Medline: 35982936]
45. Fernandez E, Salem D, Swift JK, Ramtahal N. Meta-analysis of dropout from cognitive behavioral therapy: magnitude, timing, and moderators. J Consult Clin Psychol. Dec 2015;83(6):1108-1122. [doi: 10.1037/ccp0000044] [Medline: 26302248]
46. Kumar H, Yoo S, Zavaleta Bernuy A, et al. Large language model agents for improving engagement with behavior change interventions: application to digital mindfulness. Proc ACM Hum Comput Interact. Oct 18, 2025;9(7):1-44. [doi: 10.1145/3757619]
47. Wang J, Ma W, Sun P, Zhang M, Nie JY. Understanding user experience in large language model interactions. arXiv. Preprint posted online on Jan 16, 2024. [doi: 10.48550/arXiv.2401.08329]
48. Xue F, Fu Y, Zhou W, Zheng Z, You Y. To repeat or not to repeat: insights from scaling LLM under token-crisis. Presented at: NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems; Dec 10-16, 2023; New Orleans, Louisiana, USA. [doi: 10.5555/3666122.3668712]
49. Yao J, Yang S, Xu J, Hu L, Li M, Wang D. Understanding the repeat curse in large language models from a feature perspective. In: Findings of the Association for Computational Linguistics: ACL 2025. Association for Computational Linguistics; 2025:7787-7815. [doi: 10.18653/v1/2025.findings-acl.406]

50. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol. Jan 2020;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
51. Lee YK, Suh J, Zhan H, Li JJ, Ong DC. Large language models produce responses perceived to be empathic. Presented at: 2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII); Sep 15-18, 2024; Glasgow, United Kingdom. [doi: 10.1109/ACII63134.2024.00012]
52. Tkalac Verčič A, Verčič D, Gerdina S. Situational dependence of internal communication channels. Public Relat Rev. Nov 2025;51(4):102608. [doi: 10.1016/j.pubrev.2025.102608]
53. Zhang L, Yu J, Zhang S, et al. Unveiling the impact of multi-modal interactions on user engagement: a comprehensive evaluation in AI-driven conversations. arXiv. Preprint posted online on Jun 21, 2024. [doi: 10.48550/arXiv.2406.15000]
54. Smith KA, Ward T, Lambe S, et al. Engagement and attrition in digital mental health: current challenges and potential solutions. NPJ Digit Med. Jul 2, 2025;8(1):398. [doi: 10.1038/s41746-025-01778-w] [Medline: 40604240]
55. Feng X, Tian L, Ho GWK, Yorke J, Hui V. The effectiveness of AI chatbots in alleviating mental distress and promoting health behaviors among adolescents and young adults: systematic review and meta-analysis. J Med Internet Res. Nov 26, 2025;27(1):e79850. [doi: 10.2196/79850] [Medline: 41313175]
56. Cross SP, Alvarez-Jimenez M. The digital cumulative complexity model: a framework for improving engagement in digital mental health interventions. Front Psychiatry. 2024;15:1382726. [doi: 10.3389/fpsyt.2024.1382726] [Medline: 39290300]
57. Zantvoort K. Machine learning dropout predictions for personalizing digital mental health interventions [Dissertation]. Leuphana Universität Lüneburg; 2025. [doi: 10.48548/pubdata-1596]
58. Stieglitz S, Hofeditz L, Brünker F, Ehnis C, Mirbabaie M, Ross B. Design principles for conversational agents to support Emergency Management Agencies. Int J Inf Manage. Apr 2022;63:102469. [doi: 10.1016/j.ijinfomgt.2021.102469] [Medline: 35043026]
59. Mutsuddi A, Connelly K. Text messages for encouraging physical activity are they effective after the novelty effect wears off? Presented at: 2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops; May 21-24, 2012; San Diego, United States. [doi: 10.4108/icst.pervasivehealth.2012.248715]
60. Sung J, Christensen HI, Grinter RE. Robots in the wild: understanding long-term use. Presented at: HRI '09: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction; Mar 9-13, 2009; La Jolla, CA, USA. [doi: 10.1145/1514095.1514106]
61. Prusiński T. Factors promoting alliance quality: differentiation of therapeutic alliance according to the formal aspects of the psychotherapeutic process and demographic variables. Eur J Psychiatry. Jul 2023;37(3):190-198. [doi: 10.1016/j.ejpsy.2022.11.002]
62. Bordin ES. The generalizability of the psychoanalytic concept of the working alliance. Psychother Theory Res Pract. 1979;16(3):252-260. [doi: 10.1037/h0085885]
63. Xu Z, Lee YC, Stasiak K, Warren J, Lottridge D. The digital therapeutic alliance with mental health chatbots: diary study and thematic analysis. JMIR Ment Health. Oct 10, 2025;12:e76642. [doi: 10.2196/76642] [Medline: 41072011]
64. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health. 2022;4:847991. [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
65. Darcy A, Daniels J, Salinger D, Wicks P, Robinson A. Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. JMIR Form Res. May 11, 2021;5(5):e27868. [doi: 10.2196/27868] [Medline: 33973854]
66. Hoffman V, Flom M, Mariano TY, et al. User engagement clusters of an 8-week digital mental health intervention guided by a relational agent (Woebot): exploratory study. J Med Internet Res. Oct 13, 2023;25(1):e47198. [doi: 10.2196/47198] [Medline: 37831490]
67. Ebner P, Szczuka J. Predicting romantic human-chatbot relationships: a mixed-method study on the key psychological factors. arXiv. Preprint posted online on Feb 28, 2025. [doi: 10.48550/arXiv.2503.00195]
68. Jagoe EL. Depersonalized intimacy: the cases of Sherry Turkle and Spike Jonze. Engl Stud Can. 2016;42(1-2):155-173. [doi: 10.1353/esc.2016.0004]
69. Turkle S. Alone Together - Why We Expect More from Technology and Less from Each Other. Basic Books; 2011. ISBN: 978-0-465-01021-9
70. Diebel C, Goutier M, Adam M, Benlian A. When AI-based agents are proactive: implications for competence and system satisfaction in human–AI collaboration. Bus Inf Syst Eng. Jan 3, 2026;68(2):259-278. [doi: 10.1007/s12599-024-00918-y]

71. Dwyer A, de Almeida Neto A, Estival D, Li W, Lam-Cassettari C, Antoniou M. Suitability of text-based communications for the delivery of psychological therapeutic services to rural and remote communities: scoping review. JMIR Ment Health. Feb 24, 2021;8(2):e19478. [doi: 10.2196/19478] [Medline: 33625373]
72. Islam R, Bae SW. Revolutionizing mental health support: an innovative affective mobile framework for dynamic, proactive, and context-adaptive conversational agents. arXiv. Preprint posted online on Jun 22, 2024. [doi: 10.48550/arXiv.2406.15942]
73. Kozlov E, McDarby M, Prescott M, Altman M. Assessing the care modality preferences and predictors for digital mental health treatment seekers in a technology-enabled stepped care delivery system: cross-sectional study. JMIR Form Res. Sep 15, 2021;5(9):e30162. [doi: 10.2196/30162] [Medline: 34343965]

Abbreviations

CA: conversational agent
CAI: conversational AI
CBT: cognitive behavioral therapy
EMA: ecological momentary assessment
LLM: large language model
mARM: Mobile Agnew Relationship Measure
MCAR: missing completely at random
mHealth: mobile health
Mini-SPIN: Mini–Social Phobia Inventory
NLP: natural language processing
PHQ-4: Patient Health Questionnaire–4
PHQ-D: Patient Health Questionnaire–Distress
SWLS: Satisfaction With Life Scale
UCLA: UCLA Loneliness Scale
WAI-SR: Working Alliance Inventory–Short Revised

Edited by Stephanie Law; peer-reviewed by Akshay Arora, Jinyu Guo, Tirumala Ashish Kumar Manne, Yihan Hu; submitted 16.Dec.2025; final revised version received 01.Sep.2026; accepted 01.Sep.2026; published 09.Nov.2026

Please cite as:
Schäfer LM, Krause T, Köhler S
Engagement, Feasibility, and Perceived Alliance With a Large Language Model–Enabled Mental Health Voice and Chatbot: Exploratory Longitudinal Observational Study
JMIR Form Res 2026;10:e89708
URL: https://formative.jmir.org/2026/1/e89708
doi: 10.2196/89708

© Lea Maria Schäfer, Tabea Krause, Stephan Köhler. Originally published in JMIR Formative Research (https://formative.jmir.org), 09.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.