Original Paper

AI-Enabled, Text-Based Health Coaching and Navigation for Employees to Support Health Outcomes: Pre-Post Observational Study

Paula Wilbourne[1,2*], PhD; Susan Mirch-Kretschmann[3*], PhD; Denise Walker[4], PhD; Michael Varghese[1], MBA; Roberto Arnetoli[1], MSc

[1]Sibly, Inc, San Mateo, CA, United States
[2]RecoveryWorksAI, Pacifica, CA, United States
[3]VA Palo Alto Health Care, U.S. Department of Veterans Affairs, Palo Alto, CA, United States
[4]School of Social Work, University of Washington, Seattle, WA, United States
*these authors contributed equally

Corresponding Author:

Paula Wilbourne, PhD
Sibly, Inc
400 Concar Dr.
San Mateo, CA 94402
United States
Phone: +1 801 855 6493
Email: paulawilbournephd@gmail.com

Abstract

Background: Limited, timely access to quality mental health treatment harms well-being and quality of life while costing individuals and organizations millions in increased medical spending and reduced productivity. Too few qualified professionals, inconsistent quality, and stigma thwart traditional solutions, creating the need for scalable, science-based solutions.

Objective: This report provides an overview of a novel digital health coaching service that consists of artificial intelligence (AI)–assisted, human-delivered, text-based health coaching. This report provides data evaluating the efficacy of this service for delivering mental health support, improving well-being, and enhancing workplace productivity.

Methods: This observational study analyzed operational and self-reported health data from employees of subscribing organizations who used Sibly's digital health coaching service. Data included response times, changes in expressed member sentiment, conversation topics, and adherence to motivational interviewing. A subset of members (n=38) provided pre-post self-reported assessment measures of distress, unhealthy days, and presenteeism, having engaged in at least 4 coaching conversations over a minimum of 14 days. Sentiment was evaluated using a natural language processing tool.

Results: Sibly provided quick access to interactive human coaching, with a median response time of 132 seconds. Sentiment analysis showed that 57% (878/1540) of conversations increased in positive emotions. The coaches maintained strong fidelity to the techniques of motivational interviewing, with adherence exceeding 90% (387/430). The proportion of users reporting severe distress declined from 33.3% (10/30) at baseline to 6.7% (2/30) at follow-up, representing a 79% relative reduction ($P<.001$). Participants also reported a reduction in the number of unhealthy days per month, decreasing from 19.57 to 15.87 per month ($P=.02$). Self-reported productivity improved by 18% during the study period ($P<.001$). Additionally, 61% (47/77) of users who received referrals to additional employer-sponsored benefits engaged with those resources, demonstrating effective care navigation to relevant support services.

Conclusions: This report provides an overview of novel mental health support and navigation services that use AI-enabled, text-based health coaching and care navigation. Data suggest that the services provide effective, scalable mental health support in workplace settings. The platform helped reduce distress, improve well-being, and boost productivity by offering immediate access to trained coaches and personalized guidance. These results are consistent with existing research on digital mental health services. They highlight the potential of AI-assisted coaching to improve access to care. Future research should include larger, diverse populations and more rigorous randomized controlled trials. This formative report provides data that describes and demonstrates a proof of concept for an innovative technology-enabled service that addresses the problems of scalability, access, quality, and stigma that challenge the provision of traditional mental health services.

Introduction

Mental Health Service Need

Untreated mental health problems are costly, having been linked to physical health problems, lower treatment adherence, and shorter lifespans [1,2]. Even before the pandemic, rates of emergency room visits for people with mental health diagnoses and comorbidities increased nationwide [3]. Finally, employees distressed at work impact culture, lower productivity of those around them, and burden human resource services [4].

In the years since the pandemic, the growing demand for mental health treatment has collided with the longstanding limitations in our capacity to provide it. Mental health provider shortages predated the pandemic and have worsened, with wait times averaging 5-6 weeks [5-7]. Unfortunately, patients do not respond well to treatment delays with longer wait times, resulting in greater no-show rates [8]. The pandemic has worsened mental health, straining an already burdened system. Only 20% of adults who needed emotional support received help from a mental health professional [8]. Issues such as location, timing, and provider capacity (45%); costs (39%); and not knowing how to find help (21%) prevented access to care. Other barriers included privacy concerns, not thinking their problems are serious enough, concerns about stigma, or other people finding out [9]. Traditional entry to care is designed for more severe conditions, while 76-90% of distressed populations do not want or need higher levels of care [10,11]. Evaluations of stepped-care models find that only about 20% of those seeking help needed medical treatment [11]. This mismatch wastes resources, can be off-putting to treatment seekers, and creates inefficiencies for efficacious treatment.

A large body of research identifies effective, efficient tools that hold promise in the face of the current mental health crisis. Despite strong evidence, traditional mental health and medical services have not taken robust advantage of this research. We will briefly review stepped care, the use of trained coaches and paraprofessional staff, digital interventions, text-based interventions, empirically based brief interventions, and measurement-based interventions as a foundation for the program that follows.

Quality often refers to evidence-based practices and the outcomes associated with those interventions [12]. Effective solutions to the mental health crisis must measure outcomes and ensure fidelity to empirically supported models to deliver the promised track record of efficacy. Using text-based interventions and real-time natural language processing (NLP) provides an opportunity to ensure that the services deliver quality in the form of adherence to evidence-based protocols. To be successful, quality, accessibility, and efficacy must be accomplished with fewer staff members, at a low cost, and promptly to meaningfully address the problems of existing solutions.

Stepped-care models provide lower levels of high-quality assistance, improve outcomes, and expand the capacity and accessibility of current mental health resources. Timely assistance (eg, self-help, digital tools, and coaches) can lead to lasting improvements [13-22]. Those who have to wait for assistance often continue to experience symptoms, get worse, and may never reach the same degree of improvement as those who get assistance quickly [13-18]. The advantages of self-help, digital resources, and nonmedical assistance, as initial steps in a stepped-care model, have been demonstrated in multiple studies of depression, substance use, anxiety, insomnia, and overall mental health symptoms [14-17,21-29]. Early initial steps of care prevent more severe mental health problems [13,14,23,30-33].

Trained coaches and nonprofessional staff provide another science-based option to address barriers to effective mental health treatment. Well-trained and supervised unlicensed staff can achieve similar outcomes as licensed professionals [34-36] when delivering well-specified interventions like motivational interviewing and cognitive-behavioral interventions [32,35-38]. Trained coaches and unlicensed staff have demonstrated both treatment fidelity and efficacy in the treatment of anxiety and depression [39] and in facilitating health behavior change [34,36-38]. The World Health Organization concluded that specialized staff are not required to deliver mental health intervention [40].

Fortunately, flexible options for delivering assistance, such as text and asynchronous support, go beyond traditional face-to-face or even video telehealth. SMS text messaging reduces the barriers to mental health treatment noted above, including access, location, timing, stigma, and privacy [41,42]. Digital content materials and SMS text messaging are effective tools [16,18,24,42,43]. Digital text messages written by trained coaches expand accessibility and lower the cost of mental health services with outcomes comparable to telephone coaching [39,42-49]. Text has become a ubiquitous form of communication and is the preferred method of communication for adults under 50 years of age. Over two-thirds of adults value SMS text messages from health care providers, and more than 41% are "constantly online" [50-52].

The potential of artificial intelligence (AI) in mental health care represents a revolutionary shift in managing mental health problems. Interactive and supportive chatbots make use of large language models, NLP, and machine learning (ML) to provide real-time interaction, creating a safe environment for users and offering immediate coping strategies, which may have the potential to address the shortage of mental health treatment resources [53,54]. NLP detects sentiment and language features linked to mental health symptoms. NLP can

also evaluate the quality of care and adherence to evidence-based interventions demonstrated in clinical interactions. In general, using digital interventions and AI allows for real-time optimization of mental health interventions [55,56].

One aspect of NLP, sentiment analysis, has been applied to mental health content in the published literature. Sentiment analysis measures the attitudes, sentiments, evaluations, and emotions of a speaker or writer based on the computational treatment of subjectivity in a text [55]. It is generally performed using rule-based tools and lexicons to calculate the semantic orientation of words and phrases in a text. A sentiment score between −1 and +1 indicates generally negative or positive sentiment, respectively (0=neutral sentiment). Sentiment analysis related to mental health concerns has been conducted on large text datasets, including those drawn from social media, electronic health records, narrative writing, and less frequently on smaller samples of the content of interviews or narrative writing samples [57-59].

Sibly: An Innovative Approach to Health Behavior Change

Sibly provides 24/7/365 digital coaching via text with trained human coaches to organizations that pay a fee to provide the service to their members. The platform integrates with the health and wellness resources of these organizations, allowing health coaches to navigate members to timely, effective referrals as needed during coaching. The platform gathers member and coach data to enhance service delivery.

Members maintain anonymity through display names, creating a secure channel for receiving coaching and educational materials related to health behaviors, life challenges, and mental health. The mobile-friendly platform eliminates barriers of location, waitlists, stigma, and timing, offering on-demand support.

Members use a messaging app and are assigned to a small team of health coaches, who respond collectively as "Sibly." Trained health coaches, using evidence-based practices such as motivational interviewing, mindfulness, coping skills training, and cognitive-behavioral tools, help members set goals and take action. In addition, Sibly's health coaches guide members to use self-help materials, sponsored benefits, and community resources that are recommended based on each member's individual needs.

Members receive unlimited coaching and progress reports. Baseline and follow-up well-being assessments are conducted. Personally identifiable information (ie, name, address, phone number, and date of birth) is collected and used only for emergencies (danger to self, others, or vulnerable persons) by PhD-level experts who respond in less than 30 minutes.

Protocols on the scope and development of text-based coaching proposed 5 domains: selection and training of coaches, specific coaching techniques, how to structure communication with those being coached, monitoring adherence to guidelines, and quality of coaching [39]. Sibly uses strategic recruitment and in-house, science-based coach

training to address provider shortages and scale services. Successful applicants must have a bachelor's degree and are behaviorally evaluated for empathy, coachability, and professionalism. Qualified candidates are invited to participate in a paid training program.

Training consists of 240 hours of competency-based instruction in listening skills, motivational interviewing, cognitive-behavioral tools, mindfulness, and crisis response, grounding the service in rigorous, science-based interventions. New health coaches demonstrate competence during observed training cases. To date, 98 health coaches have entered training, with 94.9% (93/98) successfully completing it. Training each coach costs approximately US $4800. Approximately 40% of these costs are offset by the services they provide to members during training.

Quality assurance ensures health coaches retain and improve their skills. Sibly health coaches participate in quarterly continuing education classes and receive monthly feedback on their coaching skills from expert trainers using the adapted Motivational Interviewing Treatment Integrity (MITI) coding system [60]. A proprietary tool using ML and AI detects adherence to training skills and provides real-time feedback. Ninety-five percent of coded quality assurance conversations meet competency guidelines similar to those specified by the MITI. Health coaches not meeting competency standards participate in weekly conversation reviews until they improve. Those unable to improve are disqualified from continuing their work as health coaches.

AI and ML play a crucial role in supporting Sibly's human health coaches and structuring communication with members. AI suggests optimal next steps, identifies service improvement opportunities, enhances engagement, and improves the member experience. ML provides the health coaches with insights in real time that might not occur to the coach, or that may be based on previous conversations. Sibly's ML tools reliably detect 40 specific topics, measure sentiment fluctuations, identify the start and end of a text conversation, and suggest optimal next steps to the health coaches. The topic models were designed to reduce bias by accurately representing topics related to race, gender, and sexual orientation that are less frequent in the population and typically overlooked in models that are not specifically trained to detect them.

The platform provides health coaches with digital notifications or "nudges." Nudges assist the health coaches by highlighting information and recommending the next steps. Nudges personalize recommendations based on member priorities and goals. For example, if a member is interested in weight loss, nudges suggest relevant content and employer-sponsored benefits related to weight loss. If a member is concerned about their mood, nudges can suggest screening measures, self-help materials, and employer-sponsored mental health benefits related to this member's concern. The health coaches refine AI nudges through real-time feedback to the model.

The text-based communication platform allows for rigorous quality assurance and data analytics. AI and ML

offer real-time feedback, reducing skill drift that can occur with episodic training. Through AI-assisted health coaching, Sibly is designed to provide personalized, high-quality, human-led coaching, resource navigation, and crisis response that is science-based, data-informed, efficient, consistent across the health coaches and, therefore, more scalable.

The current report includes observational data that describes the structure, service level, and impact of a novel digital health platform. We report on a subset of members who provided pre-post self-report questionnaires on mental and physical well-being.

Methods

Sample

We report on two data samples. First, over a 5-month period, all new members were asked to complete a baseline assessment after completion of their first coaching conversation. After 2 weeks and completion of the fourth coaching conversation, these members were asked to complete a follow-up assessment. Thirty-eight members who completed a baseline assessment went on to complete 3 additional coaching conversations and were asked to complete a follow-up questionnaire. Of the 38 eligible participants, 30 (78.9%) agreed to complete the follow-up.

All 38 members meeting these eligibility criteria were detected using ML-enabled tools that prompted the coach to send the follow-up survey within 15 minutes of the end of the fourth coaching conversation. On average, the conversations of the original 38 members completing the baseline assessment involved the exchange of 25 messages, 11 of which were from the member. There were no significant differences in the length or timing of conversations between those completing a follow-up assessment and those who did not.

Sample selection reflects a combination of both theoretical and practical considerations. Previous product analyses suggest that members are more likely to complete a questionnaire after exchanging 10 messages with a coach; therefore, the baseline symptom assessment was sent after the end of the first conversation. The follow-up questionnaire was sent after members completed their fourth conversation to allow credible exposure to the service in the eyes of our customers. We limited the window of time to 14 days to reduce the chances that members might be lost to follow-up.

Second, we report data on our larger sample of operational data for the entire population of members using the service. The number of participants available for analysis is specified in each analysis below. All data available were included. No participants were selectively omitted from these analyses.

Ethical Considerations

During registration, users confirmed that they were at least 18 years old and consented to the use of their deidentified aggregate data. This study received a retrospective exemption from the institutional review board of the University of Washington Human Subjects Division (STUDY00023309)

and was not considered human subjects research due to the observational and deidentified nature of the data collected.

All data management and analysis adhered to our terms of service and privacy policies, to which members agreed at enrollment. Messaging, operational, and survey data analyzed for ML were anonymous and processed in a separate system from member-identifying information. Data were reported only in aggregate and encrypted for privacy. Three data sources are reported here, including pre-post self-reported health data, operational data, and an ML analysis of the sentiment measured in participant conversations.

Quantitative Variables and Operational Data

Response time was measured as the seconds between a member's message and the coach's reply. Sequential member messages were counted as one.

The digital platform and phone app allowed us to measure, analyze, and optimize a number of product dimensions: response rates, time of day, sentiment, coaches per conversation, topics, and adherence to motivational interviewing. These data provide insight into the performance of the product and ways in which members used the service, which will be reported below.

Self-Reported Health-Related Questionnaire Data

Self-report measures were administered within the app. Pre-post data were then analyzed using a paired samples *t* test to determine the significance of changes made by participants using the service. Self-reported measures included the Lam Employment Absence and Productivity Scale (LEAPS) [61, 62], health-related quality of life as measured by the Center for Disease Control and Prevention's Healthy Days [63], a 2-item measure of mental anguish and distress [64,65], and a single item assessing the member's evaluation of the change in mental well-being as a result of their use of the service.

The LEAPS is a 10-item scale with documented internal consistency and external validity. It took 3-5 minutes to complete and was administered as a measure of the degree to which a participant's work performance was impacted by their mental health symptoms [60,61]. With the author's permission, distress was measured using 2 items assessing mental anguish that was first published in 2013 and later developed into a 10-item scale in 2019 [62,63]. The questions and response options are as follows: How much are you suffering emotionally (mental anguish, not pain, or discomfort in your body)? 1=absent; 2=very mild or occasionally; 3=mild, comes in moments and goes away; 4=moderate, steady and in specific moments; 5=marked, hurts all the time and does not get better; 6=severe, unbearable; 7=extreme, feels like you want to die; and 8=skip. A face-valid question assessing a member's change in mental well-being attributable to Sibly—"To what degree has your mental well-being changed as a result of working with Sibly?"—with response options that included 1=much better, at least 50% better, or improved on most days; 2=better, at least 25% better, or

improved 2-3 days per week; 3=a little better, some improvement, or better on 1 day per week; 4=no change; 5=a little worse, somewhat worse on at least 1 day per week; 6=worse, at least 25% worse, or worse 2-3 days per week; 7=much worse, at least 50% worse, or worse on most days.

Sentiment Analysis

To assess the impact of coaching on sentiment, we analyzed the change in sentiment or sentiment shift. First, within a conversation, sentiment shift was defined by the change in sentiment from the first 5 messages of the conversation to the last 5 messages of the conversation. Second, sentiment shift was measured from the start of each member's first conversation to the end of the conversation, occurring immediately prior to their completion of the follow-up questionnaire. We used VADER (Valence Aware Dictionary and Sentiment Reasoner), a sentiment analysis tool designed to process informal text, including slang, emojis, and punctuation. VADER is part of the Natural Language Toolkit [66,67], a leading suite of Python libraries and programs for NLP. A sample of our first 1512 conversations with B2B (business-to-business or participants whose membership was paid for through an employer) members was analyzed to look at the number of coaches per conversation. These reflect the total number of conversations with employee members that had taken place at the time of this analysis. No conversations or members were omitted. A chi-square test was used to look at the relationship between the length of the conversations and the sentiment of the conversation.

Coaches per Conversation

We analyzed 1549 B2B conversations to look at the number of coaches per conversation. These reflect the total number of conversations with employee members that had taken place at the time of this analysis. No conversations or members were omitted. A chi-square test was used to look at the relationship between the number of coaches per conversation (1-7) and members' experience of those conversations as reflected by the percentage of conversations with a positive sentiment shift.

Topics

We used an unsupervised topic modeling technique, Latent Dirichlet Allocation (LDA), to identify conversation topics.

LDA is a generative statistical model used in NLP and ML to identify thematic structures within large text datasets. Overall, we used 3 trained health coaches as coders to label the topics detected by the models. We then asked the coders to rate the degree to which the proposed models captured the topics discussed in 100 randomly selected conversations. The output and evaluation of this process are reported below.

Results

Response Time

Coaches responded to member messages with a median response time of 132 seconds and an average of 197 seconds.

Optimizing Questionnaire Response Rates

Before December 2021, approximately 29% (2795/9565) of our requests to members to complete surveys were acted on. To improve this response rate, we examined optimal points in the member relationship (number of messages sent) and the timing in the conversation at which members were most likely to complete a questionnaire or survey (within 15 minutes of the end of a conversation). Optimizing survey administration along these dimensions increased our response rate to 48.6% (2041/4201; χ^2_1=485.4, n=13,766, P<.001).

Time of Day and Work Hours

Fifty-six percent (8938/15,960) of our member messages were sent outside of the hours of 9 AM to 5 PM, suggesting that most employees are accessing assistance without disruption to their workday and outside the hours that most mental health services would be available to them (see Figure 1). During work hours, members take twice as long to respond to coaches as they do outside of work hours, suggesting the ability to fit coaching conversations amid other demands on their time.

Figure 1. Distribution of employee messages by time of day, 2021-2023.
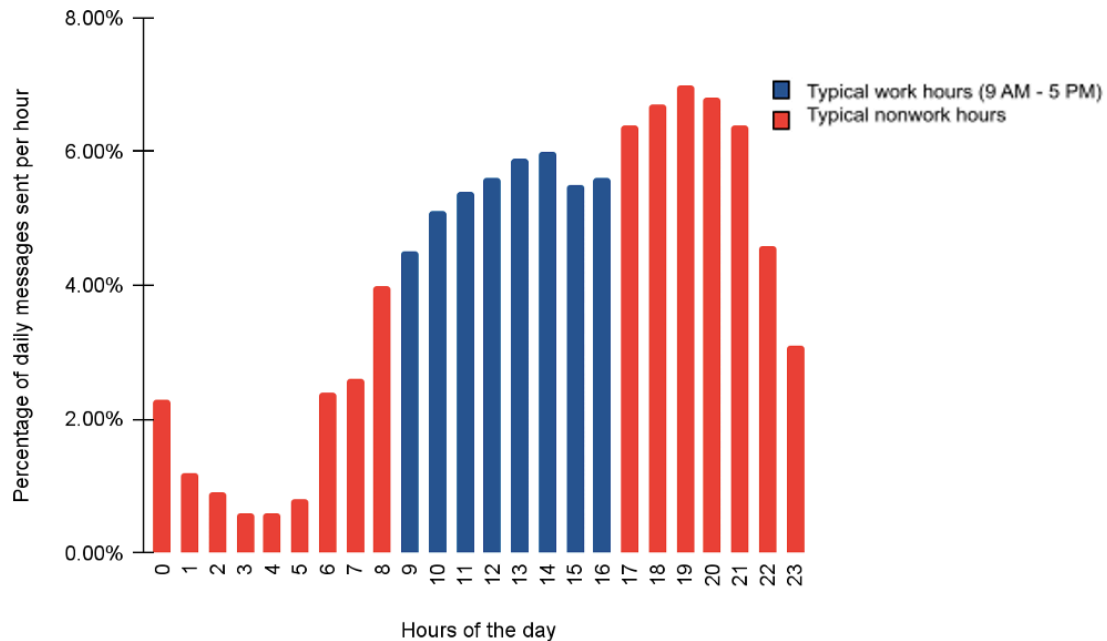
Sentiment

Over each member's relationship with Sibly, the slope of the sentiment line between the beginning of the first conversation and the end of the conversation occurring before the follow-up assessment was positive for a majority of the participants. Twenty-one (70%) of the sample of 30 members completing pre-post measures demonstrated a positive shift in sentiment, while 9 demonstrated a negative shift in sentiment. Within individual conversations, an average of 57% (878/1540) demonstrated a positive sentiment shift. This information provided insight into the impact of coaching conversations on member sentiment that occurred during the member conversations and over the initial coaching conversations. In our larger member population, there was a positive correlation between the length of member conversations with sentiment shift suggesting that the longer a member talks with a coach, the better they feel or that the better a person is feeling, the longer they are likely to continue talking to Sibly (χ^2_5=58.01, n=1512, P<.001).

Coaches per Conversation

In a larger sample of 1549 member conversations, 70% (1084/1549) of conversations were completed by a single coach, 23% (356/1549) of conversations included responses from 2 coaches, while 7% (108/1549) of conversations included 3 to 7 coaches. Sentiment analysis did not detect a decrease in sentiment associated with conversations that switched between one coach and another. In fact, conversations with more than one coach were longer and demonstrated a greater improvement in sentiment than those in which a member spoke to fewer health coaches during the conversation (χ^2_3=17.701, n=1549, P<.001). Generally, members are not aware that they have spoken to more than one coach

during a single conversation. However, in <2% (23/1549) of conversations, members express concern or dissatisfaction about working with more than one coach or the perceived change in coach during a conversation.

Topics

Using an unsupervised topic modeling technique, we developed and evaluated 7 possible models with 30, 40, or 50 topic clusters. Topics were labeled, and labels were validated by 3 independent coders. We then asked the coders to rate the fit of these models on a 5-point Likert scale for 100 randomly selected conversations. Models A-F were rated as "best reflecting the conversations" as follows: A (45/100, 45%), B (21/100, 21%), C (17/100, 17%), D (10/100, 10%), E (10/100, 10%), F (2/100, 2%), and G (1/100, 1%) of the time. Model A, with 40 topics identified, was selected as the one best reflecting the conversations 45% (45/100) of the time, with the topics detected in model A evaluated as fair to very good 84% (84/100) of the time.

Applying the selected model, we found that members came to the first conversation with an average of 5.6 topics, where over 76% (12,129/15,960) presented with at least 3 interconnected topics of conversation. Thirty-one topics of conversation were detected across all member conversations in this sample. Negative emotions, work, family, relationships, and mental health were the most common topics of conversation, occurring in more than 20% of conversations. Self-organization, self-improvement, employer-sponsored benefits, health behaviors, and love were discussed in more than 10% of conversations. Twenty-one additional topics occurred in fewer than 10% of the conversations. See Table 1 for all topics.

Table 1. Machine learning–detected topics from member-initiated messages (n=15,960).

Topic	Values, n (%)
Negative emotions	5108 (32.0)
Work	5152 (32.3)
Family	4739 (29.7)
Relationship	3783 (23.7)
Mental health	3377 (21.2)
Self-organization	2626 (16.5)
Self-improvement	2472 (15.5)
Discussing benefits	2457 (15.4)
Health behaviors	2417 (15.1)
Love	2319 (14.5)
Sleep	1385 (8.7)
Emotions: hope	1252 (7.8)
Living situation	1227 (7.7)
Hobbies	1188 (7.4)
Coping strategies	1126 (7.1)
Divorce	1115 (7.0)
Finance	1115 (7.0)
Medical problem	1090 (6.8)
Education	1050 (6.6)
COVID	1025 (6.4)
Goals focusing	875 (5.5)
Religion	666 (4.2)
Happy fun	610 (3.8)
Relaxation	548 (3.4)
Racism	337 (2.1)
Legal	272 (1.7)
LGBTQ+	265 (1.7)
Friendship	257 (1.6)
Disability	252 (1.6)
Politics	113 (0.7)
Gender	40 (0.3)

Evidence-Based Practice Adherence

During the period in which data were collected, 2 conversations per coach per month were coded for adherence to motivational interviewing using the Sibly quality assurance manual, which was based on the MITI scale [60]. This resulted in 430 coded conversations, of which 387 (90%) were determined to be adherent to motivational interviewing.

Distress

Participants were asked to rate their levels falling into categories of "1=mild or none" for those indicating "absent or occasionally," "2=moderate" for those indicating "comes and goes or steady," and "3=severe" for those indicating that it "hurts all the time, unbearable, and feels like they want to die." These groupings were chosen to aid in the interpretability when presenting to nonclinical audiences. These were referred to triage with PhD, per previously mentioned procedure, earning scores 1-3, respectively, at both baseline

and follow-up. At baseline, participants reported an average distress level of 2.27, with 33.3% (10/30) reporting severe distress. After using the service, average distress decreased to 1.87, with an 80% decrease in those reporting severe distress and a 200% increase in those reporting mild or no distress. A paired samples 2-tailed t test found a significant improvement in average distress levels, $t_{29}=-3.89$ ($P<.001$), and a medium effect size $d=0.71$.

Unhealthy Days

Participants were queried as to the number of unhealthy days in the past month attributable to mental health and then physical health. These numbers were added together up to a maximum of 30 days, consistent with the scoring recommendations for this measure by the Centers for Disease Control. Participants also reported an average of 19.57 unhealthy days per month at baseline, with 60% (18/30) experiencing poor health (16+ unhealthy days/month). At follow-up, unhealthy

days decreased to 15.87 (P=.02, d=0.42), and the proportion of members in poor health dropped to 36.7% (11/30).

Productivity or Presenteeism

Analyzing data from the LEAPS questionnaire found that at baseline assessment, participants scored an 8.5, which falls in the mild range for productivity impairment due to mental health symptoms at work. At follow-up, participants reported ~18% (1.5/8.5) decrease with a mean score of 7. A paired samples 2-tailed t test found a significant improvement in productivity, t_{29}=−3.99 (P<.001), and a medium effect size d=0.71.

Benefit Referral and Engagement

During the evaluation period, the health coaches recommended 18 benefit programs to 14 participants. Members engaged with 11 (61.1%) of the benefit services recommended to them. Additionally, among a sample of 77 participants who were referred to their EAP program, 46 of these participants enrolled in that benefit program. For comparison, a third-party comparison of Sibly to more traditional benefit navigators found that Sibly's referral success rate was 3 times higher than that of more traditional navigation services.

Discussion

This paper describes a novel AI-enabled, text-based health coaching platform that supports mental health via immediate, accessible assistance from live human coaches. Analysis found that the health coaches respond to members in less than 3 minutes. Objective quality assurance ratings found that more than 90% (387/430) of conversations were adherent to motivational interviewing and that 56% (8938/15,960) of member messages were sent outside of traditional clinical service hours of 9 AM to 5 PM. While 70% (1084/1549) of member conversations were conducted by a single coach, sentiment did not decline when multiple coaches participated; instead, conversations with multiple coaches were longer, with a greater improvement in sentiment. Over time, a majority of members showed an increase in positive sentiment between their first and last conversations and within the majority of individual conversations. The most common topics of conversation included negative emotions, work, family, relationships, and mental health. Other common topics included self-organization, self-improvement, employer benefits, health behaviors, and love. A subset of members providing pre-post self-report data reported an 80% decrease in severe distress, a 19% decrease in unhealthy days, and an 18% increase in productivity. Sixty-one percent of the time that a health coach referred a member to an additional employer-sponsored benefit, the referral was successful. The results demonstrate the potential of AI-supported text coaching as an efficient, scalable, and effective workplace mental health solution.

There are important implications for the key findings that highlight the speed of access, rapid response times, consistent use of empirically based tools, and improvements in distress, productivity, and engagement with employer-sponsored benefits. The analyses presented indicate that the service increases accessibility for individuals who may not want or cannot schedule therapy, reduces delays in care, and holds a credible promise to supplement traditional services, especially in those who need immediate support.

The service uses a one-to-many model, a novel and efficient way to scale the human coaching relationship. Each member is assigned to a small team of coaches who are selected for empathy and are trained in empirically based tools, allowing the coaches to speak with a similar voice and deliver a consistent intervention. The analysis reported shows that coach transitions within a single conversation do not impact member sentiment. This finding supports the credibility of the one-to-many model, where each member is assigned to a small team of coaches who respond as a single entity. Sibly's digital platform and AI support enable a seamless collaboration among providers, which is difficult to achieve in traditional care.

Members most frequently discussed topics of negative emotion, work, family, relationships, and mental health and aligned with previous findings that work-related stress, personal relationships, and emotional distress significantly impact mental health challenges. The fact that work stress ranked second underscores the importance of workplace mental health solutions, and these findings suggest that text-based coaching may improve mental well-being and workplace functioning. Previous research has found that mental health interventions enhance productivity and reduce absenteeism at work [68].

These findings are consistent with prior research on digital mental health interventions, which have been shown to improve access to care and support well-being in a scalable way [69]. Stepped-care models suggest that lower-intensity interventions, such as text-based coaching, can be highly effective for individuals who do not require or want more traditional therapy or clinical treatment [60]. By integrating benefit navigation, Sibly addresses a common challenge seen in both mental health support and employer-sponsored benefits—helping individuals connect with appropriate resources when they are needed. ML also played a role in increasing engagement and response rates, helping to optimize interactions in ways that traditional benefit navigation methods cannot do. This suggests that AI-assisted coaching could enhance accessibility and service quality while also supporting employee well-being and workplace productivity.

Despite these encouraging results, this study has several limitations. Because it was observational and lacked a control group, we cannot conclude with certainty that Sibly was the direct cause of the observed improvements. While members reported significant reductions in distress and unhealthy days, external factors could have influenced these outcomes. Additionally, the sample size was relatively small and composed of employees with access to employer-sponsored benefits, meaning the results may not generalize to broader populations. Furthermore, employees self-selected to use the service, which may introduce selection bias, as

those who engaged with the platform might differ systematically from those who did not. As a result, the findings may not fully reflect the experiences or needs of the larger employee population. Future research should aim to include larger and more diverse samples, as well as implement randomized controlled trials to establish causality. While sentiment analysis and engagement metrics provide useful insights into user experiences, more qualitative research is needed to understand how members perceive and engage with the platform.

The broader implications of these findings highlight the increasing role of digital health coaching in expanding access to mental health support and reducing the burden on traditional clinical services. AI-enabled platforms like Sibly offer a promising, scalable, and cost-effective solution to the growing need for workplace mental health solutions. Digital services have the potential to shift mental health care from a reactive, appointment-based medical model to a proactive, on-demand system. As AI and NLP continue to advance, these services will likely become more precise, responsive, and personalized to individual users. Additional research and transparent evaluation will be crucial in refining these models and ensuring that digital mental health solutions continue to meet the evolving needs of individuals and organizations alike.

Data Availability

Data analyzed in this study are not publicly available due to the proprietary nature of the data generated by a commercial organization.

Authors' Contributions

Conceptualization: PW (lead), SM-K (equal), DW (supporting)
Data curation: PW, MV (lead), RA
Formal analysis: MV (lead), RA (supporting), PW (supporting), DW (supporting)
Funding acquisition: PW
Investigation: PW, SM-K, DW, MV, RA
Methodology: PW, SM-K, DW, MV, RA
Project administration: DW (lead), PW (supporting)
Resources: PW
Supervision: PW
Validation: PW, SM-K, DW, MV, RA
Visualization: MV (lead), RA (supporting)
Writing – original draft: SM-K (lead), PW (equal)
Writing – review & editing: DW (lead), PW (supporting), SM-K (supporting), RA (supporting)

Conflicts of Interest

PW, MV, and RA each have a small amount of stock in the company that purchased Sibly. PW was the founder and chief science officer at Sibly, and SM-K, MV, and RA are consultants with Sibly. No other conflicts of interest exist.

References

1. Understanding the link between chronic disease and depression [NIH Publication No. 24-MH-8015]. National Institute of Mental Health. URL: https://www.nimh.nih.gov/health/publications/chronic-illness-mental-health [Accessed 2025-08-27]
2. Valiant GE. Natural history of male psychologic health – effects of mental health on physical health. N Engl J Med. 1979;301:1249-1254. [doi: 10.1056/NEJM197912063012302]
3. Capp R, Hardy R, Lindroth R, Wiler J. National trends in emergency department visits by adults with mental health disorders. J Emerg Med. Aug 2016;51(2):131-135. [doi: 10.1016/j.jemermed.2016.05.002] [Medline: 27614303]
4. Hemp P. Presenteeism: at work—but out of it. Harvard Business Review. Oct 2004. URL: https://hbr.org/2004/10/presenteeism-at-work-but-out-of-it [Accessed 2025-08-27]
5. SAMHSA. URL: https://www.samhsa.gov/about/careers/behavioral-health-workforce [Accessed 2022-05-25]
6. Gruber J, Prinstein MJ, Clark LA, et al. Mental health and clinical psychological science in the time of COVID-19: challenges, opportunities, and a call to action. Am Psychol. Apr 2021;76(3):409-426. [doi: 10.1037/amp0000707] [Medline: 32772538]

7. Dembosky A. Americans can wait many weeks to see a therapist: California law aims to fix that. NPR. Nov 18, 2021. URL: https://www.npr.org/sections/health-shots/2021/11/18/1053566020/americans-can-wait-many-weeks-to-see-a-therapist-California-law-aims-to-fix-that [Accessed 2025-08-27]

8. Gallucci G, Swartz W, Hackerman F. Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. Psychiatr Serv. Mar 2005;56(3):344-346. [doi: 10.1176/appi.ps.56.3.344]

9. Conroy J, Lin L, Ghaness A. Why people aren't getting the care they need. Monitor on Psychology. Jul 1, 2020. URL: https://www.apa.org/monitor/2020/07/datapoint-care [Accessed 2025-08-27]

10. Nicholas J, Ringland KE, Graham AK, et al. Stepping up: predictors of "stepping" within an iCBT stepped-care intervention for depression. Int J Environ Res Public Health. Nov 25, 2019;16(23):4689. [doi: 10.3390/ijerph16234689] [Medline: 31775297]

11. Richards DA, Bower P, Pagel C, et al. Delivering stepped care: an analysis of implementation in routine practice. Implement Sci. Jan 16, 2012;7(1):3. [doi: 10.1186/1748-5908-7-3] [Medline: 22248385]

12. Kilbourne AM, Beck K, Spaeth-Rublee B, et al. Measuring and improving the quality of mental health care: a global perspective. World Psychiatry. Feb 2018;17(1):30-38. [doi: 10.1002/wps.20482] [Medline: 29352529]

13. Harris KB, Miller WR. Behavioral self-control training for problem drinkers: components of efficacy. Psychol Addict Behav. 1990;4(2):82-90. [doi: 10.1037/0893-164X.4.2.82]

14. Miller WR, Baca LM. Two-year follow-up of bibliotherapy and therapist-directed controlled drinking training for problem drinkers. Behav Ther. Jun 1983;14(3):441-448. [doi: 10.1016/S0005-7894(83)80107-5]

15. Edwards G, Orford J, Egert S, et al. Alcoholism: a controlled trial of "treatment" and "advice". J Stud Alcohol. May 1977;38(5):1004-1031. [doi: 10.15288/jsa.1977.38.1004] [Medline: 881837]

16. Lancee J, van den Bout J, van Straten A, Spoormaker VI. Internet-delivered or mailed self-help treatment for insomnia? A randomized waiting-list controlled trial. Behav Res Ther. Jan 2012;50(1):22-29. [doi: 10.1016/j.brat.2011.09.012] [Medline: 22055281]

17. Schmidt MM, Miller WR. Amount of therapist contact and outcome in a multidimensional depression treatment program. Acta Psychiatr Scand. May 1983;67(5):319-332. [doi: 10.1111/j.1600-0447.1983.tb00349.x] [Medline: 6869040]

18. Moberg C, Niles A, Beermann D. Guided self-help works: randomized waitlist controlled trial of Pacifica, a mobile app integrating cognitive behavioral therapy and mindfulness for stress, anxiety, and depression. J Med Internet Res. Jun 8, 2019;21(6):e12556. [doi: 10.2196/12556] [Medline: 31199319]

19. Knapstad M, Smith ORF. Social anxiety and agoraphobia symptoms effectively treated by prompt mental health care versus TAU at 6- and 12-month follow-up: secondary analysis from a randomized controlled trial. Depress Anxiety. Mar 2021;38(3):351-360. [doi: 10.1002/da.23132] [Medline: 33393688]

20. Knapstad M, Lervik LV, Sæther SMM, Aarø LE, Smith ORF. Effectiveness of prompt mental health care, the Norwegian version of improving access to psychological therapies: a randomized controlled trial. Psychother Psychosom. 2020;89(2):90-105. [doi: 10.1159/000504453] [Medline: 31794968]

21. Edwards G, Guthrie S. A controlled trial of inpatient and outpatient treatment of alcohol dependency. Lancet. Mar 1967;289(7489):555-559. [doi: 10.1016/S0140-6736(67)92129-0]

22. Miller WR, Gribskov CJ, Mortell RL. Effectiveness of a self-control manual for problem drinkers with and without therapist contact. Int J Addict. Oct 1981;16(7):1247-1254. [doi: 10.3109/10826088109039178] [Medline: 7327785]

23. Behavioral self-control training. Williammiller.net. URL: https://williamrmiller.net/behavioral-self-control-training/ [Accessed 2025-08-27]

24. Andrews G, Cuijpers P, Craske MG, McEvoy P, Titov N. Computer therapy for the anxiety and depressive disorders is effective, acceptable and practical health care: a meta-analysis. PLoS ONE. Oct 13, 2010;5(10):e13196. [doi: 10.1371/journal.pone.0013196] [Medline: 20967242]

25. Andrews G, Tolkien II Team. Tolkien II: a needs-based, costed, stepped-care model for mental health services, clinical pathways, treatment flowcharts, costing structures. World Health Organization Collaborating Centre for Classification in Mental Health; 2006. URL: https://researchers.mq.edu.au/en/publications/tolkien-ii-a-needs-based-costed-stepped-care-model-for-mental-hea [Accessed 2025-08-27]

26. Cuijpers P, Donker T, van Straten A, Li J, Andersson G. Is guided self-help as effective as face-to-face psychotherapy for depression and anxiety disorders? A systematic review and meta-analysis of comparative outcome studies. Psychol Med. Dec 2010;40(12):1943-1957. [doi: 10.1017/S0033291710000772] [Medline: 20406528]

27. Ho FYY, Yeung WF, Ng THY, Chan CS. The efficacy and cost-effectiveness of stepped care prevention and treatment for depressive and/or anxiety disorders: a systematic review and meta-analysis. Sci Rep. Jul 5, 2016;6:29281. [doi: 10.1038/srep29281] [Medline: 27377429]

28.   Richards D, Richardson T. Computer-based psychological treatments for depression: a systematic review and meta-analysis. Clin Psychol Rev. Jun 2012;32(4):329-342. [doi: 10.1016/j.cpr.2012.02.004] [Medline: 22466510]

29.   Rivero-Santana A, Perestelo-Perez L, Alvarez-Perez Y, et al. Stepped care for the treatment of depression: a systematic review and meta-analysis. J Affect Disord. Nov 1, 2021;294:391-409. [doi: 10.1016/j.jad.2021.07.008] [Medline: 34315101]

30.   Salomonsson S, Santoft F, Lindsäter E, et al. Stepped care in primary care–guided self-help and face-to-face cognitive behavioural therapy for common mental disorders: a randomized controlled trial. Psychol Med. Jul 2018;48(10):1644-1654. [doi: 10.1017/S0033291717003129] [Medline: 29095133]

31.   van't Veer-Tazelaar PJ, van Marwijk HWJ, van Oppen P, et al. Stepped-care prevention of anxiety and depression in late life: a randomized controlled trial. Arch Gen Psychiatry. Mar 2009;66(3):297-304. [doi: 10.1001/archgenpsychiatry.2008.555] [Medline: 19255379]

32.   American Psychological Association. Guidelines for prevention in psychology. American Psychologist. 2014;69(3):285-296. [doi: 10.1037/a0034569]

33.   National Research Council and Institute of Medicine. Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities. National Academies Press; 2009. URL: https://www.fredla.org/wp-content/uploads/2022/08/Preventing-Mental-Emotional-and-Behavioral-Disorders-Among-Young-People.pdf [Accessed 2025-08-27]

34.   Diebold A, Ciolino JD, Johnson JK, Yeh C, Gollan JK, Tandon SD. Comparing fidelity outcomes of paraprofessional and professional delivery of a perinatal depression preventive intervention. Adm Policy Ment Health. Jul 2020;47(4):597-605. [doi: 10.1007/s10488-020-01022-5] [Medline: 32086657]

35.   Valentine SE, Ahles EM, Dixon De Silva LE, et al. Community-based implementation of a paraprofessional-delivered cognitive behavioral therapy program for youth involved with the criminal justice system. J Health Care Poor Underserved. 2019;30(2):841-865. [doi: 10.1353/hpu.2019.0059] [Medline: 31130554]

36.   Hattie JA, Sharpley CF, Rogers HJ. Comparative effectiveness of professional and paraprofessional helpers. Psychol Bull. May 1984;95(3):534-541. URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.909.6659&rep=rep1&type=pdf [Accessed 2025-09-25] [Medline: 6399757]

37.   Montgomery EC, Kunik ME, Wilson N, Stanley MA, Weiss B. Can paraprofessionals deliver cognitive-behavioral therapy to treat anxiety and depressive symptoms? Bull Menninger Clin. 2010;74(1):45-62. [doi: 10.1521/bumc.2010.74.1.45] [Medline: 20235623]

38.   Rubak S, Sandbaek A, Lauritzen T, Christensen B. Motivational interviewing: a systematic review and meta-analysis. Br J Gen Pract. Apr 2005;55(513):305-312. [Medline: 15826439]

39.   Lattie EG, Graham AK, Hadjistavropoulos HD, Dear BF, Titov N, Mohr DC. Guidance on defining the scope and development of text-based coaching protocols for digital mental health interventions. Digit Health. 2019;5. [doi: 10.1177/2055207619896145] [Medline: 31897306]

40.   World Health Organization. mhGAP intervention guide for mental, neurological and substance use disorders in non-specialized health settings: mental health gap action programme (mhGAP), version 2.0. 2016. URL: https://apps.who.int/iris/handle/10665/250239 [Accessed 2025-08-27]

41.   Muñoz RF, Chavira DA, Himle JA, et al. Digital apothecaries: a vision for making health care interventions accessible worldwide. Mhealth. 2018;4:18. [doi: 10.21037/mhealth.2018.05.04] [Medline: 30050914]

42.   Rathbone AL, Prescott J. The use of mobile apps and SMS messaging as physical and mental health interventions: systematic review. J Med Internet Res. Aug 24, 2017;19(8):e295. [doi: 10.2196/jmir.7740] [Medline: 28838887]

43.   Cho SMJ, Lee JH, Shim JS, et al. Effect of smartphone-based lifestyle coaching app on community-dwelling population with moderate metabolic abnormalities: randomized controlled trial. J Med Internet Res. Oct 9, 2020;22(10):e17435. [doi: 10.2196/17435] [Medline: 33034564]

44.   Markert C, Sasangohar F, Mortazavi BJ, Fields S. The use of telehealth technology to support health coaching for older adults: literature review. JMIR Hum Factors. Jan 29, 2021;8(1):e23796. [doi: 10.2196/23796] [Medline: 33512322]

45.   Lindner P, Olsson EL, Johnsson A, Dahlin M, Andersson G, Carlbring P. The impact of telephone versus e-mail therapist guidance on treatment outcomes, therapeutic alliance and treatment engagement in Internet-delivered CBT for depression: a randomised pilot trial. Internet Interv. Oct 2014;1(4):182-187. [doi: 10.1016/j.invent.2014.09.001]

46.   Gupta I, Di Eugenio B, Ziebart B, et al. Human-human health coaching via text messages: corpus, annotation, and analysis. In: Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue. Association for Computational Linguistics; 2020:246-256. [doi: 10.18653/v1/2020.sigdial-1.30]

47.   Dol J, Aston M, Grant A, McMillan D, Tomblin Murphy G, Campbell-Yeo M. Effectiveness of the "essential coaching for every mother" postpartum text message program on maternal psychosocial outcomes: a randomized controlled trial. Digit Health. 2022;8. [doi: 10.1177/20552076221107886] [Medline: 35720618]

48. Gell NM, Grover KW, Savard L, Dittus K, Mace E. Outcomes of a text message, Fitbit, and coaching intervention on physical activity maintenance among cancer survivors: a randomized control pilot trial. J Cancer Surviv. Feb 2020;14(1):80-88. [doi: 10.1007/s11764-019-00831-4] [Medline: 31776849]

49. Oreopoulos P, Petronijevic U, Logel C, Beattie G. Improving non-academic student outcomes using online and text-message coaching. J Econ Behav Organ. Mar 2020;171:342-360. [doi: 10.1016/j.jebo.2020.01.009]

50. Gelles-Watnick R. Americans' use of mobile technology and home broadband. Pew Research Center. Jan 31, 2024. URL: https://www.pewresearch.org/internet/2024/01/31/home-broadband-mobile-acknowledgments/ [Accessed 2025-08-27]

51. Rainie L, Zickuhr K. Americans' views on mobile etiquette. Pew Research Center. Aug 26, 2015. URL: https://www.pewresearch.org/internet/2015/08/26/americans-views-on-mobile-etiquette/ [Accessed 2025-08-27]

52. Campbell KJ, Blackburn BE, Erickson JA, et al. Evaluating the utility of using text messages to communicate with patients during the COVID-19 pandemic. J Am Acad Orthop Surg Glob Res Rev. Jun 15, 2021;5(6):e21.00042. [doi: 10.5435/JAAOSGlobal-D-21-00042] [Medline: 34128913]

53. Minerva F, Giubilini A. Is AI the future of mental healthcare? Topoi (Dordr). 2023;42:809-817. [doi: 10.1007/s11245-023-09932-3] [Medline: 37361723]

54. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry. Jul 2019;64(7):456-464. [doi: 10.1177/0706743719828977] [Medline: 30897957]

55. Sadeh-Sharvit S, Hollon SD. Leveraging the power of nondisruptive technologies to optimize mental health treatment: case study. JMIR Ment Health. Nov 26, 2020;7(11):e20646. [doi: 10.2196/20646] [Medline: 33242025]

56. Olawade DB, Wada OZ, Odetayo A, David-Olawade AC, Asaolu F, Eberhardt J. Enhancing mental health with artificial intelligence: current trends and future prospects. J Med Surg Public Health. Aug 2024;3:100099. [doi: 10.1016/j.glmedi.2024.100099]

57. Liu B. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers; 2012. Synthesis Lectures on Human Language Technologies. [doi: 10.1007/978-3-031-02145-9]

58. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med. Apr 8, 2022;5(1):46. [doi: 10.1038/s41746-022-00589-7] [Medline: 35396451]

59. Zunic A, Corcoran P, Spasic I. Sentiment analysis in health and well-being: systematic review. JMIR Med Inform. Jan 28, 2020;8(1):e16023. [doi: 10.2196/16023] [Medline: 32012057]

60. Moyers TB, Manuel JK, Ernst D. Motivational interviewing treatment integrity (MITI) coding manual 4.1. University of New Mexico; 2014. URL: https://motivationalinterviewing.org/sites/default/files/miti4_2.pdf [Accessed 2025-08-27]

61. Lam RW, Michalak EE, Yatham LN. A new clinical rating scale for work absence and productivity: validation in patients with major depressive disorder. BMC Psychiatry. Dec 3, 2009;9:78. [doi: 10.1186/1471-244X-9-78] [Medline: 19958540]

62. The Lam Employment Absence and Productivity Scale (LEAPS). Department of Psychiatry, University of British Columbia; URL: https://med-fom-psychiatry-wwd.sites.olt.ubc.ca/files/2012/07/LEAPS-description-and-scale.pdf [Accessed 2025-08-27]

63. Measuring healthy days: population assessment of health-related quality of life. Centers for Disease Control and Prevention; Nov 2000. URL: https://www.cdc.gov/hrqol/pdfs/mhd.pdf [Accessed 2025-08-27]

64. Tossani E. The concept of mental pain. Psychother Psychosom. 2013;82(2):67-73. [doi: 10.1159/000343003] [Medline: 23295405]

65. Fava GA, Tomba E, Brakemeier EL, et al. Mental pain as a transdiagnostic patient-reported outcome measure. Psychother Psychosom. Nov 27, 2019;88(6):341-349. [doi: 10.1159/000504024]

66. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Web Soc Media. 2014;8(1):216-225. [doi: 10.1609/icwsm.v8i1.14550]

67. NLTK. URL: https://www.nltk.org/ [Accessed 2025-8-27]

68. Burton WN, Schultz AB, Chen C, Edington DW. The association of worker productivity and mental health: a review of the literature. Int J Workplace Health Manag. Jun 27, 2008;1(2):78-94. [doi: 10.1108/17538350810893883]

69. Bond RR, Mulvenna MD, Potts C, O'Neill S, Ennis E, Torous J. Digital transformation of mental health services. npj Mental Health Res. 2023;2(1). [doi: 10.1038/s44184-023-00033-y]

Abbreviations

AI: artificial intelligence
B2B: business to business
LDA: Latent Dirichlet Allocation

LEAPS: Lam Employment Absence and Productivity Scale
MITI: Motivational Interviewing Treatment Integrity
ML: machine learning
NLP: natural language processing
VADER: Valence Aware Dictionary and Sentiment Reasoner