JMIR Mental Health

Journal Impact Factor (JIF) (2023): 4.8
Volume 13 (2026) ISSN 2368-7959 Editor in Chief: John Torous, MD, MBI

Contents

Original Papers

Reviews

Automated Safety Plan Scoring in Outpatient Mental Health Settings Using Large Language Models: Exploratory Study

Hayoung K Donnelly[1,2], PhD; Gregory K Brown[1], PhD; Kelly L Green[1], PhD; Ugurcan Vurgun[2], PhD; Sy Hwang[2], MS; Emily Schriver[2,3], MS; Michael Steinberg[1], MA; Megan E Reilly[1], MPH; Haitisha Mehta[4], MA; Christa Labouliere[4], PhD; Maria A Oquendo[1], MD, PhD; David Mandell[1], SCD; Danielle L Mowery[2,5], PhD

[1]Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, United States
[2]Institute for Biomedical Informatics, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA, United States
[3]University of Pennsylvania Health System, Philadelphia, PA, United States
[4]Department of Psychiatry, Columbia University, New York, NY, United States
[5]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, United States

Corresponding Author:
Danielle L Mowery, PhD
Institute for Biomedical Informatics, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA, United States

Abstract

Background: The safety planning intervention (SPI) is a suicide prevention intervention that results in a written plan to help patients reduce suicide risk. High-quality safety plans—that is, those that are the most complete, personalized, and specific—are more effective in reducing suicide risk. Measuring SPI quality is labor-intensive, which means that clinicians rarely get specific, actionable feedback on their use of the SPI.

Objective: This study aimed to develop the Safety Plan Fidelity Rater, an automated tool that assesses the quality of written safety plans leveraging 3 large language models (LLMs)—GPT-4, LLaMA 3, and o3-mini.

Methods: Using 266 deidentified safety plans from outpatient mental health settings in New York, LLMs analyzed four key steps: warning signs, internal coping strategies, making the environment safe, and reasons for living. We compared the predictive performance of the three LLMs, optimizing scoring systems, prompts, and parameters.

Results: Findings showed that LLaMA 3 and o3-mini outperformed GPT-4, with different step-specific scoring systems recommended based on weighted F_1-scores.

Conclusions: These findings highlight LLMs' potential to provide clinicians with timely and accurate feedback on safety plan quality, which could greatly improve its implementation in community practice.

(JMIR Ment Health 2026;13:e79010) doi:10.2196/79010

KEYWORDS
suicide; mental health informatics; generative AI; clinician support; patient-reported data; artificial intelligence

Introduction

The safety planning intervention (SPI) is a widely used, evidence-based intervention to prevent suicide [1]. It is designed to help patients develop awareness of their personal warning signs and pre-plan specific strategies to use to prevent and manage acute suicidal crises [1]. The written safety plan that results from SPI has six steps: (1) identifying internal warning signs of an impending acute suicidal crisis; (2) identifying distracting activities the individual can do by themselves; (3) identifying people and social settings that can be used for distraction; (4) identifying friends and family that the individual can reach out to for help; (5) identifying professionals and professional resources the individual can reach out to for help; and (6) developing an action plan to reduce access to lethal means. In some cases, the written safety plan includes an optional Step 7, which identifies the patient's reasons for living. SPI reduces suicidal behavior following emergency department discharge by almost half compared to usual care [2]. Unfortunately, evidence suggests that clinicians do not implement SPI as intended, which likely decreases its effectiveness in reducing suicidal behavior [3-5].

Providing real-time feedback to clinicians could greatly enhance their use of the SPI [4,6,7]; however, the current state of the science requires direct observation by expert clinicians to provide this feedback. Even providing useful, asynchronous feedback based on recorded SPI sessions is challenging. While an alternative approach of rating the quality of the written safety plan is more scalable, using human coders still requires significant time and training [8].

Applying large language models (LLMs) in mental health care services can improve service delivery [9]. Existing LLM applications have primarily focused on risk prediction, rather than on evaluating the quality of clinical care. One study found that only 15% of LLM applications in mental health research focused on supporting treatment and intervention practices, whereas 47% concentrated on risk prediction [10]. Even among the few studies on treatment and intervention support, most have focused on clinical documentation summarization and information extraction, rather than providing feedback on treatment quality. There has been some attempt to use LLMs to improve suicide prevention. For example, one study evaluated the presence of the components of the SPI documented in the electronic health record using traditional rule-based natural language processing methods, finding that computational approaches can be used to understand suicide prevention interventions in real-world settings [11]. This effort could be further advanced by applying LLM techniques to assess the fidelity of written safety plans in a more nuanced and contextually sensitive manner without the need for ontological reasoning.

In this pilot study, we developed and pilot-tested an automated Safety Plan Fidelity Rater (SPFR), an artificial intelligence (AI) aid for scoring safety plans. We developed the SPFR by comparing three LLMs based on their weighted F_1-scores and optimizing their prompts and parameters to maximize performance. We hypothesized that o3-mini would outperform GPT-4 and LLaMA 3 in effectively scoring written safety plans compared to scores assigned by trained clinical coders because of its advanced reasoning capabilities.

Methods

Data

We analyzed 266 deidentified written safety plans collected as part of a Zero Suicide implementation project [12]. Safety plans were collected from patients who received the SPI across 61 outpatient mental health clinics in New York State. Patient demographic data were not available in this study; only the written responses on the safety plans were used for analysis. Written responses on the safety plan are intended to be collaboratively developed by the clinician and the patient together. Coders assessed safety plan response quality using the SPI Scoring Algorithm (SPISA; GKB, unpublished data, 2023). The SPISA consists of 20 items measuring the quality and completeness of written responses across seven steps of the safety plan form: warning signs, internal coping strategies, social contacts and social settings, social contacts for a crisis, professionals for a crisis, making the environment safe, and reasons for living. The quality of response codes evaluates how specific and personalized the response is, as well as its relevance to each step's purpose. Completeness of response codes evaluates whether a response is present or absent on a given line. We focused on developing an automated scoring tool for the quality of responses, particularly for the following four steps of the Safety Plan: warning signs, internal coping strategies, making the environment safe, and reasons for living. The remaining three steps were excluded due to a high number of missing entries following the deidentification process, as they often contained personal information (eg, names and phone numbers) that were redacted in compliance with the Health Insurance Portability and Accountability Act of 1996 using the Safe Harbor method.

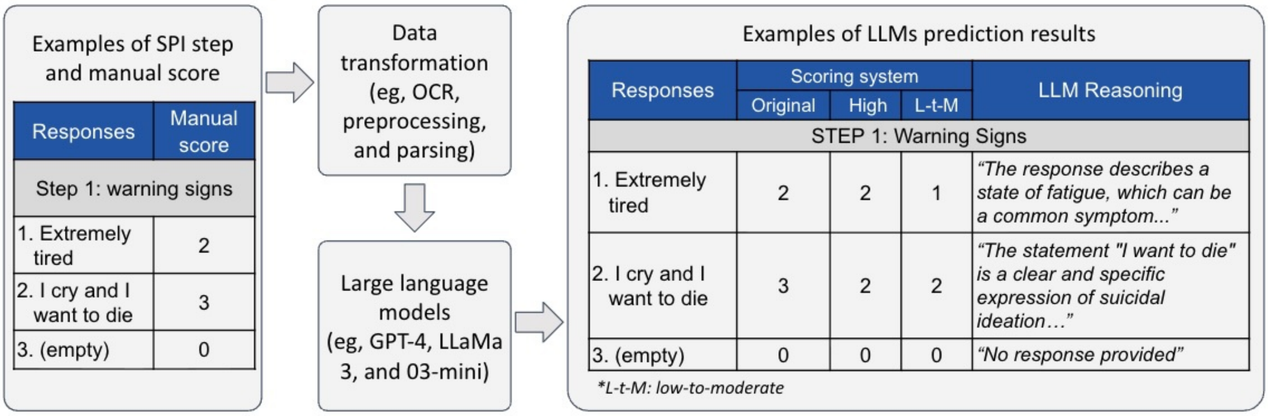
Below are the definitions of the four target steps, along with fictitious but realistic examples:

- Warning signs*:* specific thoughts, feelings, physiological states, or behaviors that are associated with the development of a suicidal crisis (eg, "not sleeping well," "my heart starts pounding faster").
- Internal coping strategies*:* activities the patient can engage in by themselves that distract from suicidal urges and allow time for the crisis to pass (eg, "write in my journal," "listening to the Beatles").
- Making the environment safe*:* an action plan that the client will complete to reduce access to lethal means (eg, "get rid of unused medication," "lock up my firearm and give the key to my brother").
- Reasons for living: things that matter most to clients and give them a sense of purpose and motivation to continue living (eg, "my family," "I want to go to college, become a nurse, and help people").

The maximum possible number of responses differed for each step: 3 for warning signs, 3 for internal coping strategies, 2 for making the environment safe, and 1 for reasons for living. If more than the maximum number of responses were present for a given step, the SPISA dictates for the highest coded responses to be chosen up to the maximum number (eg, if for warning signs, 4 responses were given of 3, 3, 2, and 1, then the top 3 highest scores of 3, 3, and 2 are chosen). Among 2210 individual responses across four steps from the 266 safety plan forms, we observed 772 responses for warning signs, 770 responses for internal coping strategies, 405 responses for making the environment safe, and 263 responses for reasons for living.

In Figure 1, we present a workflow diagram illustrating an example of the safety plan steps, the SPFR development, and output.

Figure 1. Workflow of the Safety Plan Fidelity Rater development. LLM: large language model; OCR: optical character recognition; SPI: safety planning intervention.



Optical Character Recognition

The original data files are PDF documents containing typed safety plans without handwritten content. We applied Tesseract [13], an open-source text recognition engine, to extract text for LLM development. After converting the PDF files to TIFF format using the Python Imaging Library (Python Software Foundation), we applied Tesseract version 5 with default configurations for English text extraction, generating plaintext output.

Parsing

The original dataset included safety plan responses along with metadata (eg, step titles, operational definitions, and page numbers). First, to focus solely on responses, metadata elements were removed. Second, documents with clear "Step X" headings (eg, "Step 1. Warning signs") were parsed directly to extract the response content for each step. In cases without explicit step markers, pattern matching based on recurring labels (eg, "Warning signs: 1., 2.… Internal Coping Strategies: 1., 2.…") was used to segment the responses. Third, for the final step (eg, "Step 7: Reasons for living"), which lacks a clear ending marker (such as a "Step 8"), we used other textual indicators appearing after the last step (eg, "The Stanley-Brown Safety Plan...") to identify the end of the response. Finally, after extracting the responses for each step, they were further processed because each response requires separate scoring. If the responses were numbered (eg, "1. 2. 3."), this numbering was used to split and identify each response individually. In instances where the responses were not numbered, the text was divided by new line characters, and each line was treated as a distinct response.

Prompt Development and Scoring Algorithm

We used a zero-shot prompting strategy to rate each response and compared the model performance of three scoring systems. The prompt was developed based on the SPISA coding manual and iteratively refined to enhance the model's performance throughout the process. We tested 3 scoring systems: the original scoring system based on SPISA, a high-precision system, and a low-to-moderate precision system. The 4-score original system consists of four levels: (0) no response, (1) general, (2) somewhat specific, and (3) highly specific. Examples of responses with varying levels of precise specificity and corresponding scores are provided in Textbox 1. The 3-score high precision system has (0) no response, (1) general, and (2) somewhat and highly specific. The 3-score low-to-moderate precision system has (0) no response, (1) general and somewhat specific, and (2) highly specific responses. To improve the explainability of the rater [14], we also incorporated reasoning generation for each score, ensuring that clinicians receive interpretable feedback.

Textbox 1. Prompt example for step 1: warning signs.

Instruction Prompt = '''Given the following responses, evaluate the quality of written personal "Warning signs among patients with suicide risk. Warning signs are specific thoughts, feelings, physiological states, or behaviors that may indicate suicidal crisis. For each response, return a score based on the criteria and provide clear reasoning for the score.

Criteria for scoring:

0 - No response provided, responses stating "Nothing provided" or "Nothing listed," or responses that do not clearly indicate that the clinician at least tried to elicit a response (e.g., "Client was very guarded", "Client declined to answer").

1 - Not relevant to warning signs of suicide (e.g., "Call therapist"), vague thoughts (e.g., "Bad thoughts"), unexplained emotions (e.g., "Moody", "Angry", "Sad", "Frustrated", "Stressed"), or unclear situations (e.g., "Relationship problems").

2 - Somewhat personalized. The response includes some specific details of the thoughts (e.g., "Repeated thoughts", "Don't want to talk to anyone"), feelings (e.g., "Start to get anxious", "Not in a good mood"), and situations (e.g., "Not sleeping well", "I get headaches", "Having argument with husband") but lacks enough detail to fully assess suicide risk.

3 - Highly personalized. The responses include specific thoughts (e.g., "Any suicidal thoughts, plans, or intentions," "When I think I can't take it anymore," "Thoughts like I want to be left alone, leave me alone"), detailed feelings (e.g., "Feeling trapped and stuck," "Feeling an overwhelming emptiness," "The feeling of having no one to talk to"), or intense symptoms (e.g., "Clenched fists," "Hearing voices", "Excessive sleep disturbance").

Combination responses: If a response combines multiple elements where each element would independently score a 2 (e.g., "Start to get anxious and start crying"), assign a score of 3 instead of 2. If a response combines multiple elements where each element would independently score a 1, assign a score of 1.

Response: {response_text}

Output:

Score: [score]

Reasoning: [Clearly outline the specific indicators used for assessment and explain how they are scored. Provide concrete examples or threshold criteria to justify your decision, making the prediction results more explainable]'''

Large Language Models

Using Microsoft Azure Databricks, we assessed the performance of the Safety Plan Fidelity Rater (SPFR) tool using the SPISA-coded data and three LLMs: GPT-4, LLaMA 3, and o3-mini. We originally selected only GPT-4 because it tends to perform clinical coding tasks well, but we wanted to understand whether an open-weight model, such as LLaMA 3, was comparable in performance. In addition, o3-mini's reasoning ability outperformed this baseline model. For GPT-4 and LLaMA 3, we compared models using 3 temperature settings (0.1, 0.5, and 0.9). For o3-mini, we evaluated performance across three reasoning effort levels (low, medium, and high). Four steps of the safety plan were evaluated by comparing the weighted F_1-score (a measure of how accurately a model makes predictions by balancing the impact of different types of prediction errors and ensuring that categories with more data have a greater influence on the final score, with scores ranging from 0 to 100). In this study, we define the best-performing model as the one with the highest weighted F_1-score, indicating the most accurate predictive performance. We selected F_1-score because it is a recommended metric especially for multiclass prediction with imbalanced data [15], providing a more reliable measure of model performance by balancing both precision and recall. We additionally conducted a McNemar test to examine whether the best-performing model is significantly better statistically than the other LLMs.

In Textbox 1, we provide an example prompt for Step 1: warning signs using the original 4-value SPISA scoring system and Table 1 for the associated predictions and reasoning for each of the 3 LLMs.

Table . Examples of prediction results with reasoning for Step 1: warning signs for each large language model.

Large language model	Prediction results
GPT-4	• Response: ‘excessive sleeping, hearing voices’ • Rater’s score: 3 • Large language model prediction score: 3 • Reasoning: The response includes highly personalized and intense symptoms that are specific and clearly indicative of a suicidal crisis. “Excessive sleeping” and “hearing voices” are detailed and severe indicators that highlight significant distress and potential danger, both of which are crucial for assessing suicide risk. These elements align with the criteria for a score of 3, which encompasses specific and highly personalized thoughts, feelings, or symptoms.
LLaMA 3	• Response: ‘excessive sleeping, hearing voices’ • Rater’s score: 3 • Large language model prediction score: 3 • Reasoning: The response includes two specific indicators that are highly relevant to warning signs of suicide.
o3-mini	• Response: ‘excessive sleeping, hearing voices’ • Rater’s score: 3 • Large language model prediction score: 3 • Reasoning: The response includes two specific and detailed indicators. “Excessive sleeping” is a concrete behavioral or physiological sign, and “hearing voices” is an intense symptom that falls under highly personalized indicators. Both elements are explicitly mentioned in the criteria as examples of intense or specific warning signs, making the overall response highly personalized. These details go beyond vague or general statements and clearly indicate significant risk factors, justifying a score of 3.

Ethical Considerations

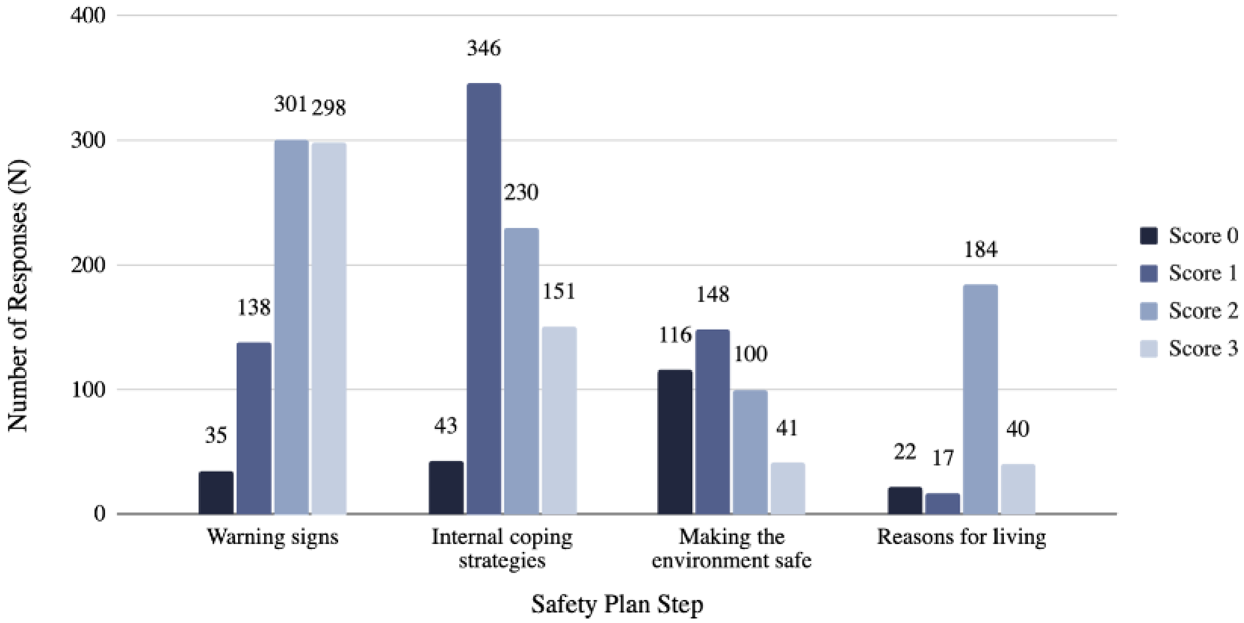
This study was approved by the University of Pennsylvania Institutional Review Board (852245) with the data use agreement and transfer (68141/02) from Columbia University.

Results

Data Characteristics

The number of responses scored by coders, with scores of 0, 1, 2, or 3, displayed an asymmetrical distribution (Figure 2). For warning signs responses (n=772), the highest proportion of responses (301/772, 38.99%) received a score of 2, followed by 38.6% (298/772) scoring 3, 17.88% (138/772) scoring 1, and 4.53% (35/772) scoring 0. For internal coping strategies (n=770), 44.94% (346/770) of responses scored 1, followed by 29.87% (230/770) scoring 2, 19.61% (151/770) scoring 3, and 5.58% (43/770) scoring 0. Among responses on making the environment safe (n=405), 36.54% (148/405) scored 1, followed by 28.64% (116/405) scoring 0, 24.69% (100/405) scoring 2, and 10.12% (41/405) scoring 3. Finally, for reasons for living (n=263), most responses (69.96%, 184/263) scored 2, with 15.21% (40/263) scoring 3, 8.37% (22/263) scoring 0, and 6.46% (17/263) scoring 1 (Figure 2).

Figure 2. Distribution of responses by quality scores coded by trained raters.



The prediction model for the quality scores of safety plan responses varied according to different scoring systems and LLMs (Table 2). We assessed four key questions: (1) How well do LLMs perform using the original scoring system? (2) Do LLMs have improved performance with high and low-to-moderate precision scoring systems? (3) Does one LLM outperform the others more consistently across steps? and (4) Which LLM is most consistent in its ratings?

Table . Weighted F_1-score across large language models. Mean refers to the average weighted F_1-score across different hyperparameters (eg, temperature for GPT-4 and LLaMA 3, reasoning effort for o3-mini).

SPI[a] step	GPT-4					LLaMA 3					o3-mini				
	Mean	SD	Min	Max	Range	Mean	SD	Min	Max	Range	Mean	SD	Min	Max	Range
Step 1: warning signs															
Original	48.93	0.27	48.57	49.23	0.66	51.78	0.31	51.47	52.22	0.75	38.12	1.70	36.66	40.30	3.64
High	73.51	1.07	72.11	74.71	2.60	78.68	0.37	78.45	79.12[b]	0.67	59.53	0.79	59.03	60.46	1.43
L-t-M[c]	63.76	3.01	63.29	64.39	1.10	65.39	0.38	65.11	65.94	0.83	63.04	1.10	61.84	64.02	2.18
Step 2: internal coping strategies															
Original	53.27	0.59	52.70	54.08	1.38	60.19	0.54	59.59	60.65	1.06	58.86	1.84	56.56	60.73	4.17
High	64.22	0.54	63.52	64.85	1.33	72.77	0.55	72.11	73.18	1.07	70.50	1.63	68.67	72.57	3.90
L-t-M	76.92	0.77	76.08	77.94	1.86	79.33	0.17	79.13	79.53	0.40	80.15	1.65	78.30	81.13[b]	2.83
Step 6: making environment safe															
Original	56.84	0.38	56.32	57.24	0.92	58.52	0.14	58.35	58.68	0.33	59.73	0.87	58.52	60.52	2.00
High	64.33	0.33	63.92	64.72	0.80	67.38	0.19	67.16	67.61	0.45	67.90	0.72	66.86	68.55	1.69
L-t-M	75.01	1.17	73.61	75.73	2.12	74.40	0.56	73.74	74.82	1.08	74.38	1.96	72.96	76.56[b]	3.60
Step 7: reasons for living															
Original	75.79	1.68	73.65	77.22	3.57	78.86	0.58	78.38	79.52	1.14	74.47	4.55	69.26	78.81	9.55
High	88.23	2.07	85.67	89.97	4.30	90.69	0.27	90.38	91.02	0.64	91.72	0.24	91.42	91.99[b]	0.57
L-t-M	81.61	1.73	79.52	82.71	3.19	84.21	0.09	84.15	84.34	0.19	80.56	4.06	75.39	84.44	9.05

[a]SPI: safety planning intervention.

[b]The best-performing model for this step.

[c]L-t-M: low-to-moderate precision system scale.

How well do LLMs score the original scoring system? The mean performance (F_1-score) for original scoring systems ranged by steps: warning signs: 38.12 - 51.78; internal coping strategies: 53.27 - 60.19; making environment safe: 56.84 - 59.73; and reasons for living: 74.47 - 78.86.

Do LLMs have improved performance with high and low-to-moderate precision scoring systems? We observed increases in predictive power with augmented precision scoring systems. The most elevated mean performance by scoring systems ranged by steps: warning signs (high): 59.53 - 78.68; internal coping strategies (low-to-moderate): 76.92 - 80.15; making environment safe (low-to-moderate): 74.38 - 75.01; and reasons for living (high): 88.23 - 91.72.

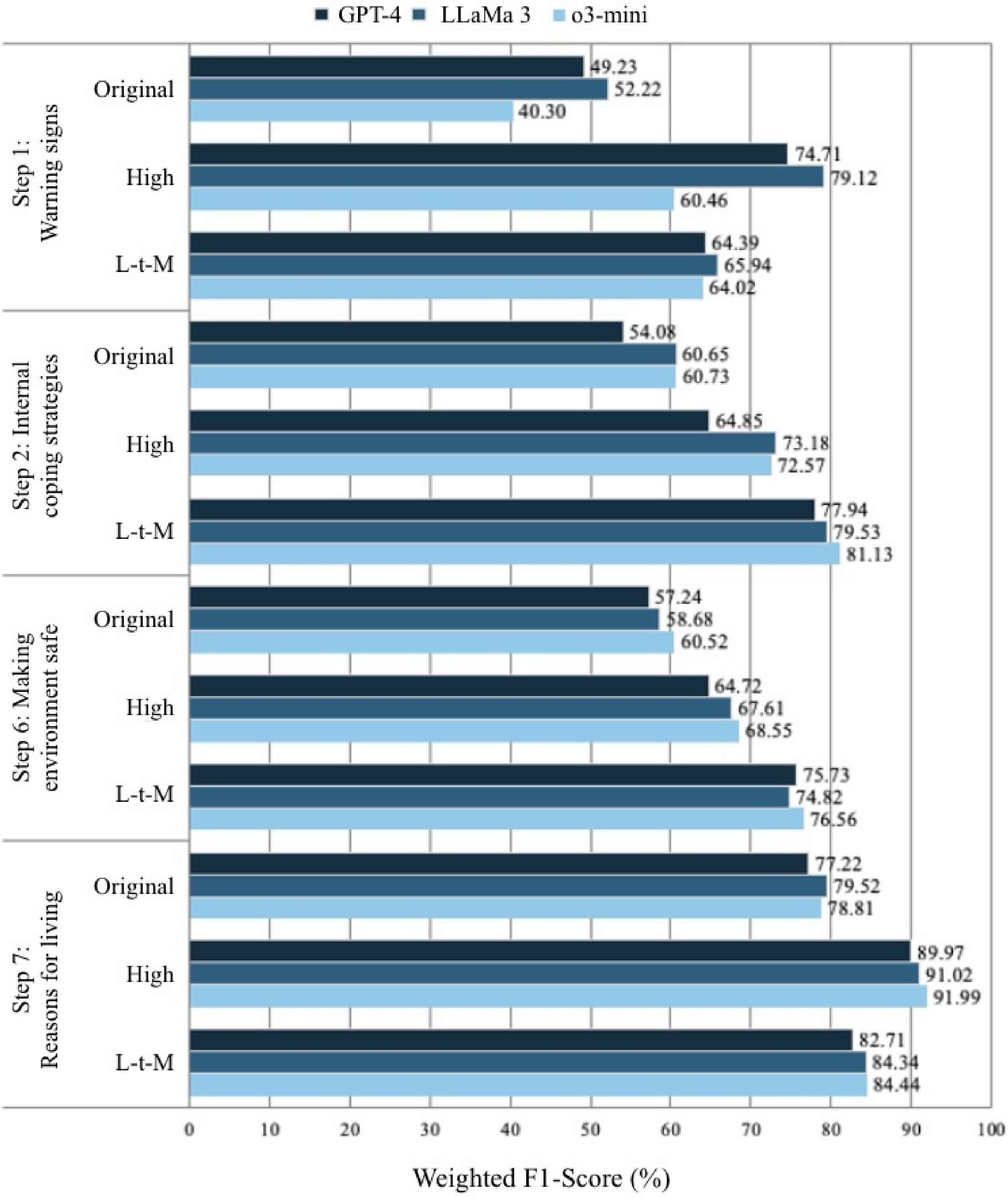
Does one LLM outperform the others more consistently across steps? LLaMA 3 produced the best predictive performance with the highest F_1-score for warning signs (high): 79.12; o3-mini produced the best predictive performance for internal coping strategies (low-to-moderate): 81.13, making environment safe (low-to-moderate): 76.56, and reasons for living (high): 91.99.

Which LLM is most consistent in its ratings? LLaMA 3 (0.72) produced the most consistent performance, with the smallest mean range difference across different parameter values, compared to GPT-4 (1.99) and o3-mini (3.72).

Best-Performing Model for Each Step

The best-performing model for each step, defined as the model with the highest weighted F_1-score across LLMs and scoring systems, was as follows (Figure 3). In addition, we conducted a McNemar test to examine whether the best-performing model, based on the highest weighted F_1-score, differed significantly from the other LLMs. In other words, we compared the model with the highest weighted F_1-score against each of the other models; for example, GPT-4 versus LLaMA 3, LLaMA 3 versus o3-mini, and o3-mini versus GPT-4.

Figure 3. Best-performing model performance across steps, scoring systems, and large language models. L-t-M: low-to-moderate precision system scale.



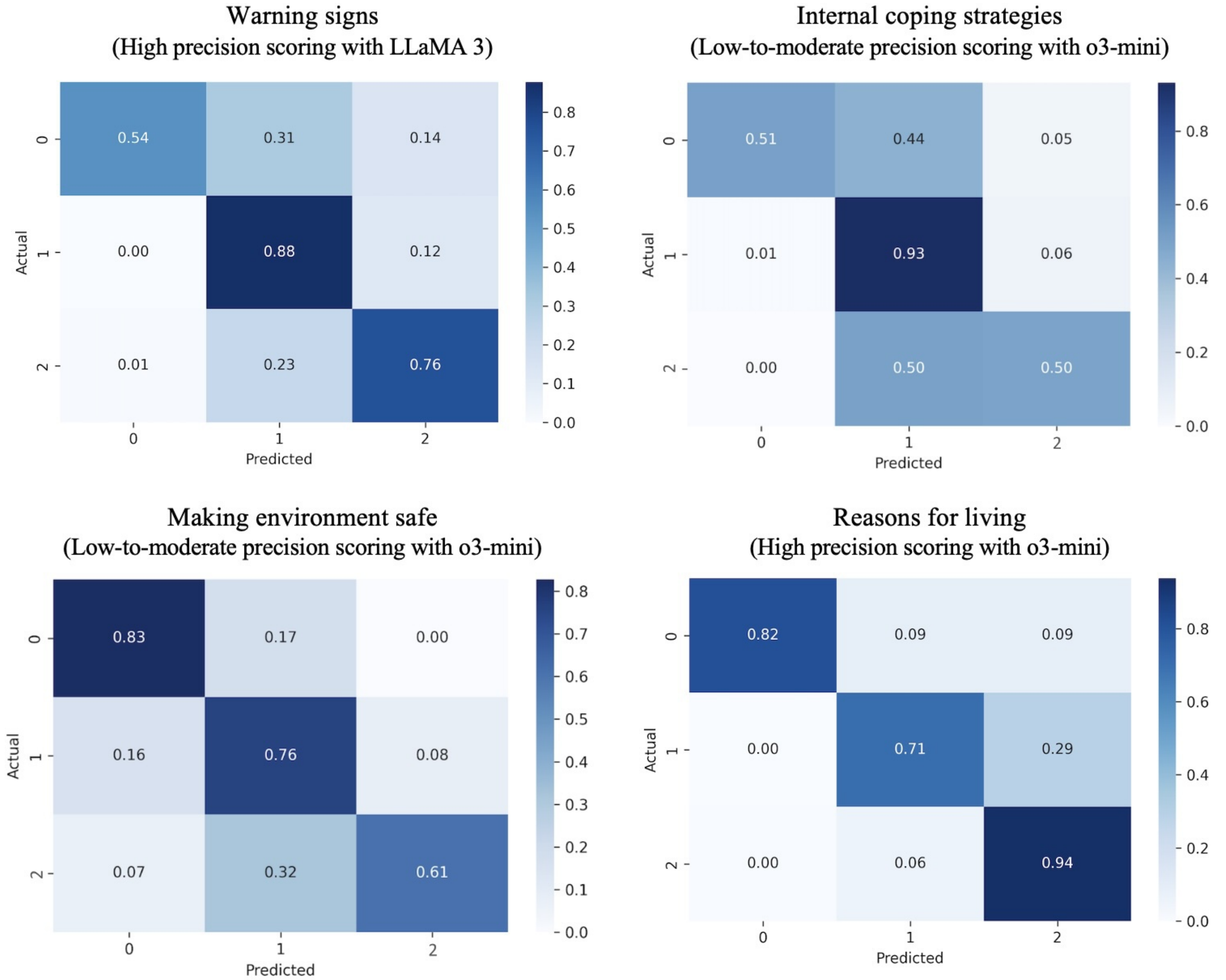
- Step 1: Warning signs: The best-performing system was the high precision scoring system using LLaMA 3 with a temperature of 0.1, achieving a weighted F_1-score of 79.12%. According to the McNemar test result, this LLaMA 3-based model showed a statistically significantly higher F_1-score than both GPT-4 (*P*=.01) and o3-mini (*P*<.001).
- Step 2: Internal coping strategies: The best-performing system was the low-to-moderate precision scoring system using o3-mini with medium reasoning effort, achieving a weighted F_1-score of 81.13%. This model showed a statistically significantly higher F_1-score than GPT-4 (*P*<.001), but no significant difference compared to LLaMA 3 (*P*=.25).
- Step 6: Making the environment safe: The best-performing system was the low-to-moderate precision scoring system using o3-mini with medium reasoning effort, achieving a

weighted F_1-score of 76.56%. Although this model reported the highest weighted F_1-score among the 3 LLMs, there were no statistically significant differences in F_1-score compared to GPT-4 (P=.08) or LLaMA 3 (P=.39).

- Step 7: Reasons for living: The best-performing system was the high precision scoring system using o3-mini with medium reasoning effort, achieving a weighted F_1-score of 91.99%. This model reported no statistically significant differences in F_1-score compared to GPT-4 (0.87) and LLaMa 3 (P=.10).

The confusion matrix in Figure 4 illustrates how well the top-performing models' predicted scores align with the raters' scores across different categories. For Step 1, warning signs responses with the high precision scoring system, among those scored as 0 by coders, 54% (19/35) were correctly predicted as 0 by the LLM model, while 31% (11/35) were misclassified as 1, and 14% (5/35) were misclassified as 2. Among responses scored as 1, 88% (121/138) were correctly predicted as 1, while 12% (17/138) were misclassified as 2. Among responses scored as 2, 76% (455/599) were correctly predicted as 2, while 1% (6/599) were misclassified as 0, and 23% (138/599) were misclassified as 1.

Figure 4. Confusion matrix of the best-performing large language model.



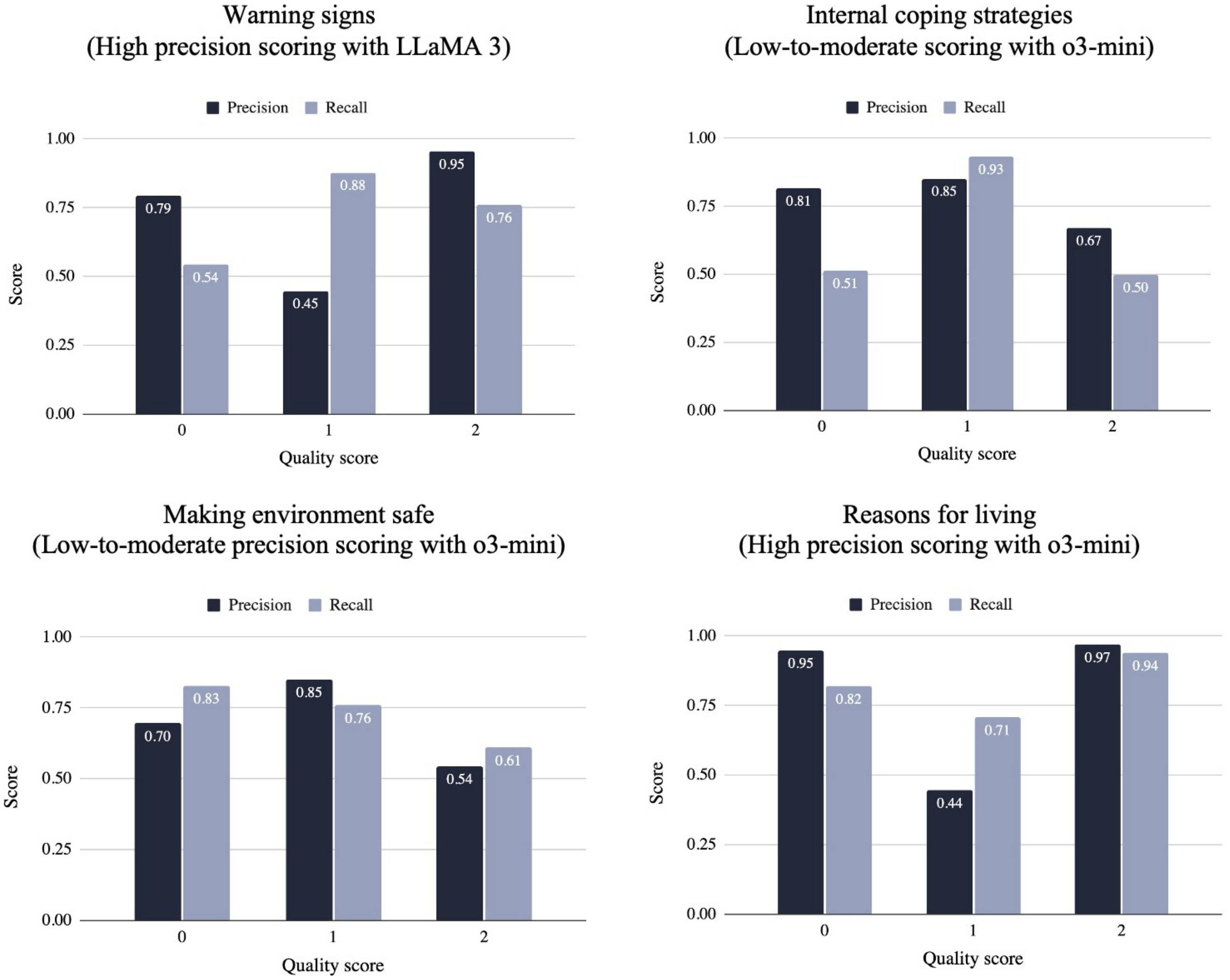
For Step 2, internal coping strategies responses with the low-to-moderate precision scoring system, among those scored as 0 by coders, 51% (22/43) were correctly predicted as 0 by the LLM model, while 44% (19/43) were misclassified as 1, and 5% (2/43) were misclassified as 2. Among responses scored as 1, 93% (536/576) were correctly predicted as 1, while 1% (6/576) were misclassified as 0, and 6% (35/576) were misclassified as 2. Among the responses rated as 2, 50% (76/151) were correctly predicted as 2, while 50% (76/151) were misclassified as 1.

For Step 6, making the environment safe responses with the low-to-moderate precision scoring system, among those scored as 0 by coders, 83% (96/116) were correctly predicted as 0 by the LLM model, while 17% (20/116) were misclassified as 1. Among responses scored as 1, 76% (188/248) were correctly predicted as 1, while 16% (40/248) were misclassified as 0, and 8% (20/248) were misclassified as 2. Among responses scored as 2, 61% (25/41) were correctly predicted as 2, 7% (3/41) were misclassified as 0, and 32% (13/41) were misclassified as 1.

For Step 7, reasons for living responses with the high precision scoring system, among those scored as 0 by coders, 82% (18/22) were correctly predicted as 0 by the LLM model, while 9% (2/22) were misclassified as 1, and another 9% (2/22) were misclassified as 2. Among responses scored as 1 by coders, 71% (12/17) were correctly predicted as 1, while 29% (5/17) were misclassified as 2. Among responses scored as 2, 94% (211/224) were correctly predicted as 2, and 6% (13/224) were misclassified as 1.

The results across different steps indicate variations in precision and recall for each score value (Figure 5). In Step 1, warning signs, Score 0 demonstrated moderate precision (0.79) and low recall (0.54); Score 1 had low precision (0.45) yet high recall (0.88); and Score 2 had high precision (0.95) with moderate recall (0.76). In Step 2, internal coping strategies, Score 0 achieved high precision (0.81) but had low recall (0.51); Score 1 showed both high precision (0.85) and recall (0.93); and Score 2 showed both low precision (0.67) and recall (0.50). In Step 6, making the environment safe, Score 0 had a moderate precision (0.70) and high recall (0.83); Score 1 had high precision (0.85) and moderate recall (0.76); and Score 2 had both low precision (0.54) and recall (0.61). In Step 7, reason for living, Score 0 had both high precision (0.95) and recall (0.82); Score 1 had low precision (0.44) but moderate recall (0.71); Score 2 had both high precision (0.97) and recall (0.94).

Figure 5. Precision and recall of the best-performing large language model.



Discussion

Principal Findings

We assessed the performance of 3 LLMs for scoring written safety plans. SPFR accuracy improved when using the 3-point scoring systems compared to the original 4-point scoring system. No one LLM provided the most optimal performance across steps and scoring systems. The findings of this study offer significant methodological advancements and areas for future research, particularly as they apply to the clinical implications of this line of research.

From a clinical practice perspective, while existing LLM applications in suicide prevention focus on screening, diagnosis, or delivering eHealth services to patients [9,16], to our knowledge, this study is innovative by demonstrating the utility of LLM in assessing written safety plan quality. At this stage, this work is too premature to apply different scoring systems practically, and further research is needed to determine the best LLM and scoring system to deploy across all steps of the safety plan. Future research will also evaluate the associations of the different scoring systems with patient outcomes (eg, suicidal behavior) and determine if changing the original scoring system is useful. A deeper understanding of different LLM-based scoring systems across all the steps of the safety plan and their clinical implications is essential for optimizing the provision of reliable, accurate feedback to clinicians. Specifically, given the preliminary nature of this work, further research is needed to optimize and select the best LLM model for scoring the entire written safety plan, as this pilot work only focused on four of the seven steps. In addition, before modifying the original scoring system to improve rating performance, research is needed to understand how potential modifications in scoring might impact associations with patient outcomes (eg, suicidal behavior) to determine if changes to the scoring system are both face valid and warranted. For instance, we can test the hypothesis that the SPFR implemented with the high precision scoring system predicts patient suicidal behavior with greater accuracy than the SPFR implemented with the low-to-moderate scoring systems. Beyond predictive accuracy and associations with patient outcomes, future research should also explore different implementation strategies for providing feedback to clinicians in order to design a tool that is most useful for clinical practice. The ultimate goal of this line of research is to establish that the automatic scoring system that is designed can enhance the quality of written safety plans. Incorporating qualitative evaluations, such as experts' agreement with LLMs' reasoning, can further improve the interpretability and acceptability of AI-generated feedback. Furthermore, embedding this tool within electronic health records systems would enable direct integration with documented safety plans, improving intervention quality by providing timely and actionable feedback to clinicians. For example, an integrated SPFR in the electronic health records that automatically rates safety plans and provides pop-up feedback to the clinician in real time may lead to increases in the quality of patient safety plans, which, in turn, may result in further reducing suicide risk for those patients who receive this intervention.

From a methodological perspective, this study is a first step in a line of inquiry about engineering questions to consider when designing an automatic scoring tool, such as the SPFR, using medical record data. At this phase, the reported LLM models remain under experimental testing, and future work is needed to evaluate and improve their clinical utility. Beyond selection of which LLMs to assess, how to set the scoring systems, and which are most consistent in their ratings, other potential considerations include: (1) selecting one LLM model for all 7 steps of safety plan or select the best-performing model for each step, (2) creating an LLM ensemble-based voting approach to assign a single score for each step, and (3) introducing few-shot learning and emerging methodologies to optimize LLM performance. Further, ethical considerations will be essential for the use of this tool in clinical practice. For instance, improving the transparency of LLM predictions can help clinicians understand the rationale behind the scores better, thereby supporting informed and responsible decision-making. In addition, evaluating these tools using data from more diverse and generalizable samples will be important for reducing bias and promoting the fairness of AI.

This study has several limitations. First, we assessed a selected sample, which consisted of typed safety plans that closely aligned with the Stanley-Brown Safety Plan form. Hence, these findings may not generalize to settings where safety plans are handwritten or vary in formatting. Our post hoc evaluation using the McNemar test revealed that performance differences between the LLMs (GPT-4, LLaMa 3, and o3-mini) were only partially significant. Therefore, it may be premature to draw definitive conclusions about which LLM performs best. Further evaluation with larger and more generalizable datasets is recommended.

Related Works

This study is one of the few, but critical, emerging works in automated methods for scoring, characterizing, and assessing the efficacy of written safety plans. Boggs et al [11] developed a natural language processing and rules-based system based on the ConText algorithm for identifying documented professional contacts, lethal means counseling for firearms, and lethal means counseling for medication access and storage from safety plans. Our study builds upon this work as the first study to apply LLMs to automatically score the quality of safety plans.

Conclusions

From this pilot project, we conclude that LLMs have the potential to support an automatic SPFR system and have identified clear paths toward improving LLM scoring performance and SPFR methodological development.

Funding

This study was supported by the National Institute of Mental Health P50 MH127511 (GKB and MO, MPIs) and R01 MH112139 (Dr Barbara H Stanley, principal investigator).

Conflicts of Interest

MO receives royalties from the Research Foundation for Mental Hygiene for the commercial use of the Columbia Suicide Severity Rating Scale. She serves as an advisor to Mind Medicine (pro bono), and Fundación Jiménez Díaz. She reviews grants for Alkermes and her family formerly owned stock in Bristol Myers Squibb (sold March 2024).

References

1. Stanley B, Brown GK. Safety planning intervention: a brief intervention to mitigate suicide risk. Cogn Behav Pract 2012 May;19(2):256-264. [doi: 10.1016/j.cbpra.2011.01.001]
2. Stanley B, Brown GK, Brenner LA, et al. Comparison of the safety planning intervention with follow-up vs usual care of suicidal patients treated in the emergency department. JAMA Psychiatry 2018 Sep 1;75(9):894-900. [doi: 10.1001/jamapsychiatry.2018.1776] [Medline: 29998307]

3. Green JD, Kearns JC, Rosen RC, Keane TM, Marx BP. Evaluating the effectiveness of safety plans for military veterans: do safety plans tailored to veteran characteristics decrease suicide risk? Behav Ther 2018 Nov;49(6):931-938. [doi: 10.1016/j.beth.2017.11.005] [Medline: 30316491]
4. Kearns JC, Crasta D, Spitzer EG, et al. Evaluating the effectiveness of safety plans for mitigating suicide risk in two samples of psychiatrically hospitalized military veterans. Behav Ther 2025 Mar;56(2):438-451. [doi: 10.1016/j.beth.2024.08.001] [Medline: 40010911]
5. Gamarra JM, Luciano MT, Gradus JL, Wiltsey Stirman S. Assessing variability and implementation fidelity of suicide prevention safety planning in a regional VA healthcare system. Crisis 2015;36(6):433-439. [doi: 10.1027/0227-5910/a000345] [Medline: 26648231]
6. Brown GK, Batdorf WH, Dedert EA, et al. National implementation of advanced training in the safety planning intervention in the Department of Veterans Affairs health care system. Psychol Serv 2025 Aug;22(3):465-476. [doi: 10.1037/ser0000880] [Medline: 38900568]
7. Boudreaux ED, Larkin C, Vallejo Sefair A, et al. Effect of an emergency department process improvement package on suicide prevention: the ED-SAFE 2 cluster randomized clinical trial. JAMA Psychiatry 2023 Jul 1;80(7):665-674. [doi: 10.1001/jamapsychiatry.2023.1304] [Medline: 37195676]
8. Beidas RS, Maclean JC, Fishman J, et al. A randomized trial to identify accurate and cost-effective fidelity measurement methods for cognitive-behavioral therapy: project FACTS study protocol. BMC Psychiatry 2016 Sep 15;16(1):323. [doi: 10.1186/s12888-016-1034-z] [Medline: 27633780]
9. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health 2024 Oct 18;11(1):e57400. [doi: 10.2196/57400] [Medline: 39423368]
10. Jin Y, Liu J, Li P, et al. The applications of large language models in mental health: scoping review. J Med Internet Res 2025 May 5;27:e69284. [doi: 10.2196/69284] [Medline: 40324177]
11. Boggs JM, Yarborough BJH, Clarke G, et al. Development and validation of electronic health record measures of safety planning practices as part of zero suicide implementation. Arch Suicide Res 2025;29(3):654-667. [doi: 10.1080/13811118.2024.2394676] [Medline: 39193908]
12. Stanley B, Labouliere CD, Brown GK, et al. Zero suicide implementation-effectiveness trial study protocol in outpatient behavioral health using the A-I-M suicide prevention model. Contemp Clin Trials 2021 Jan;100:106224. [doi: 10.1016/j.cct.2020.106224] [Medline: 33220488]
13. Smith R. Tesseract OCR engine (version 5) [software]. GitHub. 2021. URL: https://github.com/tesseract-ocr/tesseract [accessed 2025-10-23]
14. Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S. Towards interpretable mental health analysis with large language models. arXiv. Preprint posted online on Oct 11, 2023. [doi: 10.48550/arXiv.2304.03347]
15. Riyanto S, Sitanggang IS, Djatna T, Atikah TD. Comparative analysis using various performance metrics in imbalanced data for multi-class text classification. IJACSA 2023;14(6). [doi: 10.14569/IJACSA.2023.01406116]
16. Hua Y, Liu F, Yang K, et al. Large language models in mental health care: a scoping review. arXiv. Preprint posted online on Jan 1, 2024. [doi: 10.2196/preprints.64088]

Abbreviations

AI: artificial intelligence
LLM: large language model
SPFR: Safety Plan Fidelity Rater
SPI: safety planning intervention
SPISA: Safety Planning Intervention Scoring Algorithm

Edited by J Torous; submitted 13.Jun.2025; peer-reviewed by M Gasmi, M Okonkwo, P Onyekwelu; revised version received 31.Aug.2025; accepted 03.Sep.2025; published 08.Jan.2026.

Please cite as:
Donnelly HK, Brown GK, Green KL, Vurgun U, Hwang S, Schriver E, Steinberg M, Reilly ME, Mehta H, Labouliere C, Oquendo MA, Mandell D, Mowery DL
Automated Safety Plan Scoring in Outpatient Mental Health Settings Using Large Language Models: Exploratory Study
JMIR Ment Health 2026;13:e79010
URL: https://mental.jmir.org/2026/1/e79010
doi:10.2196/79010

Original Paper

Effect of a Pragmatic eHealth Behavioral Gestational Weight Gain Intervention on Household Chaos in Pregnant People of Lower Socioeconomic Status: Randomized Controlled Trial

Chelsea L Kracht[1], PhD; Kaja Falkenhain[2], PhD; Emily W Flanagan[2], PhD; Abby D Altazan[2], MS; Hannah E Cabre[2], PhD; Maryam Kebbe[3], PhD; Emily K Woolf[2], PhD; Robbie Beyl[2], PhD; Daniel S Hsia[2], MD; John W Apolzan[2], PhD; Leanne Redman[2], PhD

[1]Department of Internal Medicine, University of Kansas Medical Center, Kansas City, KS, United States
[2]Clinical Sciences Division, Pennington Biomedical Research Center, Baton Rouge, LA, United States
[3]University of New Brunswick, Fredericton, Canada

Corresponding Author:
Leanne Redman, PhD
Clinical Sciences Division
Pennington Biomedical Research Center
6400 Perkins Road
Baton Rouge, LA, 70808
United States
Phone: 1 225 763 0947
Email: leanne.redman@pbrc.edu

Abstract

Background: Household chaos is an emerging risk factor for childhood obesity development, especially in families with lower socioeconomic status (SES). It is unclear if changes in household chaos, especially in pregnancy, may mediate the effectiveness of weight-related behavioral interventions.

Objective: This study aimed to describe how household chaos changed across gestation and determine whether household chaos mediated the effect of an eHealth behavioral gestational weight gain (GWG) intervention in pregnant people with low SES.

Methods: Pregnant people who were enrolled in the US Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) were recruited for a randomized controlled trial testing the effectiveness of an eHealth-based pragmatic intervention for GWG management. The usual care group received the standard WIC program guidance and monthly health coach support with general pregnancy recommendations. The intervention group received the standard WIC program plus health information via email and weekly health coach discussions to promote healthy eating and adequate physical activity. Weight and household chaos were measured at baseline (early pregnancy, 10^{+0} to 16^{+6} weeks gestation) and at the end of the intervention (late pregnancy, 35^{+0} to 37^{+6} weeks gestation). Household chaos changes across time were examined using a paired *t* test for the continuous score and using the McNemar test for household chaos category (improved or no change vs declined). Serial linear regression models and mediation analyses assessed the relationship between the intervention group (predictor), household chaos change (mediator), and GWG (outcome) with adjustment for covariates.

Results: Among 258 participants, 53.9% (n=139) were Black, 43.4% (n=112) were nulliparous, 36.0% (n=93) were obese, and almost half (n=124, 48.1%) were classified as low household chaos at baseline. Overall, there were minimal changes in household chaos scores from early to late pregnancy (P=.34), although scores and categories tended to be higher in late pregnancy. Household chaos changes were divided; some improved or had no change (n=140, 54.3%), and some declined (n=118, 45.7%) across gestation. Household chaos did not mediate the effect of the intervention on GWG.

Conclusions: In this sample, household chaos did not change across gestation and did not explain the effect of an eHealth behavioral GWG intervention in pregnant people with lower SES. Routine-focused and multilevel interventions may improve upon these findings to support an organized home for future parent and child health.

Trial Registration: ClinicalTrials.gov NCT04028843; https://www.clinicaltrials.gov/study/NCT04028843

(JMIR Ment Health 2026;13:e74146) doi:10.2196/74146

KEYWORDS

pregnancy; stress; disadvantaged; depression; anxiety

Introduction

The early pregnancy environment can shape the development and future health of a child [1]. Maternal stress, either acute or chronic, may negatively impact the intrauterine environment through increasing allostatic load, which leads to physical and neurological consequences [2]. Maternal stress, coupled with poor mental health, has established negative implications for adverse pregnancy outcomes [3], poor child cognition [4], and overweight and obesity in adult offspring [5]. Individuals from lower socioeconomic statuses (SES) may have a heightened risk for prenatal stress due to continuous exposure to known stressors [6] such as economic and housing instability [7]. Innovative approaches to lessen maternal stress and improve mental health among populations facing economic disadvantages may help improve long-term maternal and child health.

Interventions during pregnancy to improve maternal mental health have mixed or null results, with some promising interventions using mindfulness approaches [8]. However, when considering stress, two separate systematic reviews of 41-44 studies documented that eHealth interventions were effective at improving pregnant people's stress [9], including people with lower SES [10]. Lifestyle interventions for gestational weight gain (GWG) management have the potential to aid in establishing routine and healthy lifestyle behaviors and, in turn, reduce stress. A GWG intervention that trained health professionals improved anxiety in 205 pregnant people with obesity [11]. These results are likely attributable to the intervention's focus on both diet and exercise, as a different randomized controlled trial providing supervised exercise training only did not improve maternal mental health [12]. Adopting a multibehavior lifestyle intervention for appropriate GWG to eHealth modalities may improve upon existing effective GWG interventions and reduce maternal stress, including in people with lower SES who are traditionally hard to reach, underserved, and have limited access to resources.

Household chaos is an established factor for poor child development and childhood obesity [13]. Household chaos differs from individual stress, as it evaluates stress at the home level; it is characterized by crowding, disorder, and noise in the home [14]. Higher levels of household chaos may negatively impact pregnant people's ability to sustain lifestyle behaviors and manage stress, such as practicing mindfulness, having adequate physical activity, or creating a routine within the day [15]. Making individual changes such as reducing screen time and prioritizing healthy lifestyle behaviors (eg, sleep and family meals) may improve household chaos as demonstrated in a randomized pilot study [16]. These healthy lifestyle behavior improvements align with common elements of GWG interventions. However, a systematic scoping review of 111 studies found no studies examining household chaos that were conducted in pregnant people [13]. Household chaos literature has primarily focused on parents with young children [13,15,16], and existing evaluations in pregnancy have used household chaos as a covariate [17], identifying a significant knowledge gap. There is potential for household chaos to increase across gestation as the mother prepares for birth, and this increase may be offset by prenatal interventions targeting individual health behaviors to improve maternal and child health. Accordingly, an eHealth multicomponent lifestyle intervention to promote recommended GWG was delivered in pregnant people enrolled in the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) [18,19]. There was no significant difference in the incidence of appropriate GWG according to prepregnancy BMI-specific National Academy of Medicine 2009 guidelines. However, the intervention group demonstrated lower study-observed and weekly GWG as well as lower deviation from GWG guidelines relative to the control group. This paper describes the results of a preplanned, but not preregistered, secondary analysis using data collected from the trial [19], thereby filling a gap in the literature by examining household chaos across gestation in the context of an evidence-based lifestyle intervention. Accordingly, this study seeks to (1) describe changes in household chaos and stress-related constructs across pregnancy in a lower SES population and (2) examine if changes in household chaos mediate the effect of an eHealth intervention on appropriate GWG in a lower SES population. We hypothesized that household chaos would increase across gestation (aim 1); further, household chaos changes would negatively impact and mediate the effect of a lifestyle intervention on GWG (aim 2).

Methods

Participants and Procedures

Pregnant people enrolled in WIC within the US state of Louisiana were recruited during 2019-2023 for a 2-arm parallel design randomized controlled trial aimed at increasing adherence to GWG recommendations [20]. In brief, women were eligible to enroll if they had a singleton viable pregnancy, were WIC recipients for their current pregnancy, were less than 16 weeks gestational age, had a BMI between 18.5 and 40.0 kg/m^2, had access to a smartphone with internet access, and were willing to be identified on social media to other participants. Exclusion criteria were ages younger than 18 years or older than 40 years, current drug use (including tobacco and alcohol), non–pregnancy-related chronic disease (cancer, heart disease, HIV, or type 1 or type 2 diabetes), hypertension (systolic blood pressure >160 mm Hg or diastolic blood pressure >110 mm Hg at screening), current unstable mental health or an eating disorder, plan to move out of the state less than 1 year post partum, or inability or unwillingness to complete a run-in period task of keeping an activity diary with at least 80% compliance. Eligible participants were randomized (stratified by their respective state region and BMI category) to either a behavioral intervention for appropriate GWG (alongside usual WIC services; intervention group) or to receive usual WIC services only (usual care group). The trial was designed to test the hypothesis that individuals in the intervention group would have a higher incidence of appropriate GWG (as defined by the National Academy of Medicine) relative to the usual care group

[21]. Details of the study protocol are further described elsewhere [19].

This secondary analysis used data collected at the baseline visit in early pregnancy (10^{+0} to 16^{+6} weeks gestation) and at the end-of-intervention visit in late pregnancy (35^{+0} to 37^{+6} weeks gestation). At baseline, participants completed a demographic questionnaire and other questionnaires related to mental health, and paper forms were entered into a secure online platform [22]. At the end-of-intervention visit, participants repeated the mental health questionnaires in the same manner. Anthropometrics were assessed in-person at both visits.

Intervention Groups

In brief, the behavioral intervention ("Healthy Beginnings") was an approximately 24-week eHealth intervention focused on self-monitoring of weight and weight-related health behaviors and included personalized feedback from a trained health coach [19]. The intervention consisted of weekly lessons in the form of short videos and content related to adequate physical activity, healthy eating habits, and self-monitoring of weight based on evidence-based practices. The lessons were supplemented with weekly individual coaching check-ins, a closed Facebook group to interact with other study participants, and rewards for engaging in the intervention (eg, watching videos and self-monitoring weight). Rewards could be redeemed for pregnancy and postpartum-related items (eg, diapers). Participants also received a cellular-enabled scale and a Fitbit to promote self-monitoring and data from which coaches used as additional tools for individual counseling sessions. Topics specific to routine and stress included lessons on time management, meal preparation, behavior chains, and building social support, which all occurred in the first 8 lessons, while later lessons (week 17-24) included information on mindfulness and relaxation techniques, stress and sleep, and postpartum depression (Table S1 in Multimedia Appendix 1). No major changes to eligibility, version content, bugs, or content occurred during the study.

The usual care group received standard WIC services, which included general weight management advice, and received monthly check-ins with a health coach to encourage study retention. Closed Facebook groups included pregnancy-related topics other than physical exercise and healthy eating (ie, non–weight-related).

Household Chaos

The Confusion, Hubbub, and Order Scale was used to assess household chaos in early and late pregnancy and was previously validated in mothers of young children [14]. This 15-item questionnaire investigates agreement with statements related to the participant's current household, including disorder, crowding, and noise. Questions include four Likert scale response options, ranging from "strongly disagree" to "strongly agree," with reverse coding for 8 questions. Responses were summed, and total scores ranged from 15 to 60, with a higher score indicating a more chaotic home. The total score was further categorized based on previous investigations of household chaos into 4 categories: low (score <25), moderate to low (score 25-30), moderate to high (score 31-35), and high (score >35) [23]. Change in household chaos across gestation was calculated by subtracting early pregnancy scores from late pregnancy scores.

Gestational Weight Gain

Height and weight were measured during study visits by study staff before randomization in early pregnancy, and weight was measured again in late pregnancy. Height was measured using a portable stadiometer, and weight was measured twice using a standardized, calibrated electronic scale (Tanita Corp) to the nearest 0.1 kg with the participant wearing light clothing. BMI was calculated using the standard formula (kg/m^2). GWG was calculated by subtracting early pregnancy weight from late pregnancy weight.

Covariates

We explored potential covariates for inclusion in statistical models based on past literature on maternal stress and household chaos, which included demographics, other mental health constructs, and gestational diabetes status [11,12]. At baseline, participants completed a questionnaire on their race or ethnicity, marital status, and parity. Participants also completed the 21-item Depression Anxiety Stress Scales (DASS-21), a questionnaire containing statements related to subscales of depression, anxiety, and stress [24]. Participants rated how much a statement applied to them over the prior week with 4 response options ranging from "never" to "always." Subscale questions were summed, with a higher score indicating more frequent symptoms in that subscale. Consistent with another study of prenatal stress and household chaos [17], baseline stress was included as a covariate in our statistical models to isolate the effect of household chaos. Gestational diabetes status was obtained through abstraction from birth certificate records received from the state following delivery in late pregnancy. Sleep duration was measured using an accelerometer (ActiGraph GT3x+) placed on the nondominant wrist and worn for 7 days during early and late pregnancy. Overnight sleep time was identified using GGIR version 3.0.5 [25]; this R package uses an algorithm that identifies sleep based on raw accelerometer data [26,27].

Statistical Analysis

Participants in both the intervention group and the usual care group who had data for early and late pregnancy household chaos, GWG, and covariates were retained for analysis. Frequencies and means of baseline characteristics were calculated for the entire sample and compared by intervention group using chi-square or independent-sample *t* tests for categorical and continuous data, respectively.

For the first aim, a paired *t* test was used to determine changes in household chaos across gestation using the entire sample. We also explored changes in depression, anxiety, and stress as measured via the DASS-21 [24]. Household chaos category changes across time were compared using a McNemar test. Household chaos change score was categorized as improvement or no change (score change ≤0) or worsening (score change >0). The no change group was combined with the improvement group, as this may be considered another positive indicator of household chaos, since no changes were seen across a stressful time (ie, pregnancy). Identified covariates were compared between these change groups. To further describe the sample and changes across pregnancy, we also examined within-group changes among BMI categories (normal weight, overweight, obesity) and treatment groups using paired *t* tests. Within-group comparisons of early pregnancy, late pregnancy, and change in household chaos variables with weight-related variables were examined using Pearson correlations.

For the second aim, the mediation model of intervention (predictor), household chaos change (mediator), and GWG (outcome) were tested. Given that the primary paper demonstrates an intervention effect on GWG [18], this analysis was not presented. Covariates that were associated with baseline household chaos (P<.20) were retained, including race and ethnicity (non-Hispanic White, non-Hispanic Black, Hispanic, and mixed or other), parity (categories 0, 1, 2+), BMI category at randomization, time of enrollment with regard to COVID-19 (before March 2020, March 2020-March 2021, April 2021, and after), and gestational diabetes status (yes or no; Table S2 in Multimedia Appendix 1). Time of enrollment registered slightly above our threshold, but was retained due to existing literature of higher chaos in low-income families during this period [28]. Baseline household chaos was significantly correlated with all 3 subscales of the DASS-21 (Table S3 in Multimedia Appendix 1), and the stress scale was chosen as a covariate to align with past investigations of household chaos and stress [15]. For mediation, we conducted a linear regression model of calculated household chaos change with the intervention's fixed effect and adjustment for early pregnancy stress, race and ethnicity, parity, BMI category, time of enrollment, gestational diabetes status, and early pregnancy household chaos. Then we conducted two more linear regression models: (1) the calculated GWG with the fixed effect of the intervention group, and (2) the calculated household chaos changes with the outcome of the calculated GWG using a similar approach and adjustment for the same covariates. Those models retained the same predictor, outcome, and covariates for adjustment. This approach was conducted with the PROCESS vs3.5 macro with 10,000 bootstrap intervals with unstandardized estimates [29]. Analyses were conducted using R statistical software (R Foundation for Statistical Computing) and SAS (version 9.4; SAS Institute), and statistical significance was set at P less than .05.

Ethical Considerations

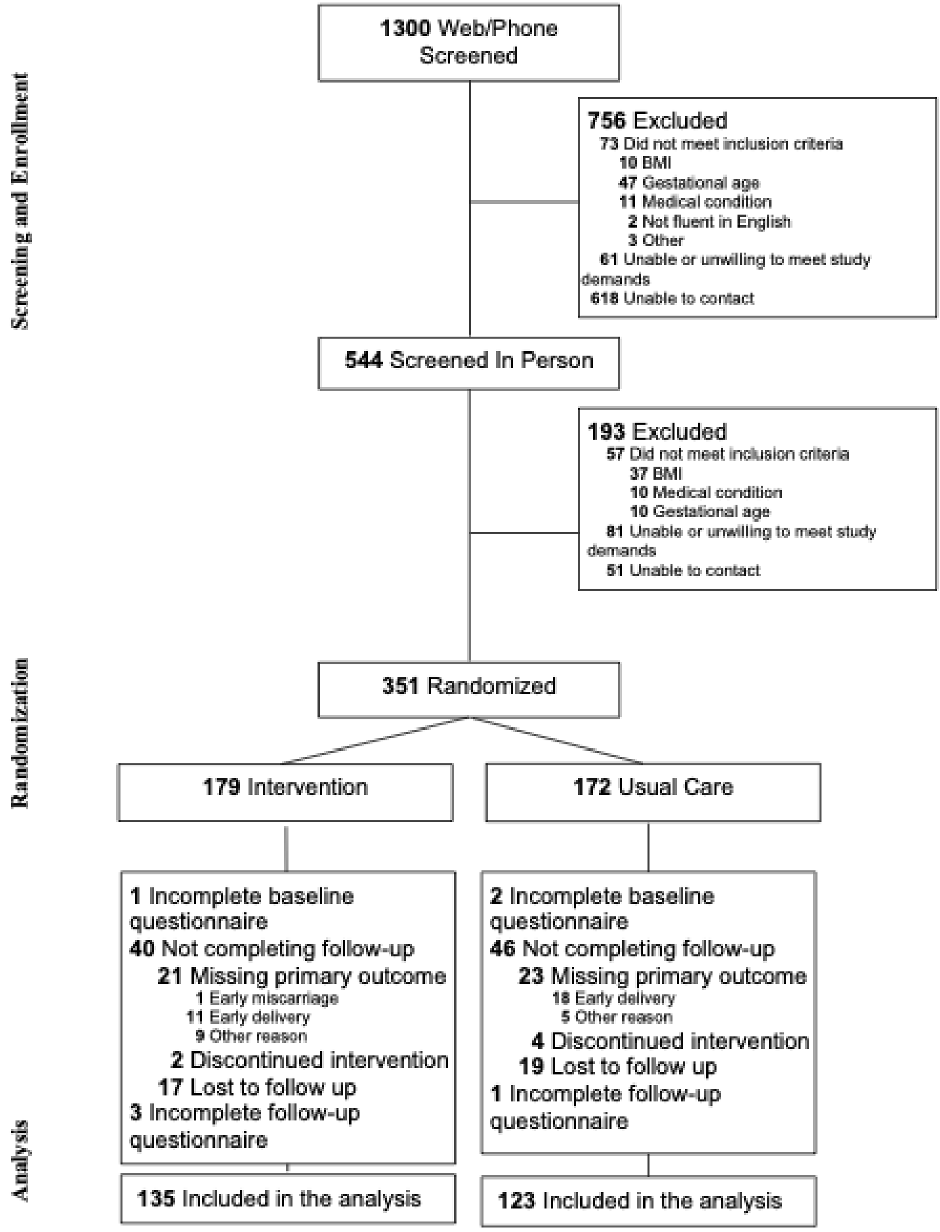
Pennington Biomedical Research Center provided institutional review board and ethics approval (2018-039-PBRC). Written informed consent was obtained at the early pregnancy visit, and from all participants in the study prior to study procedures. The informed consent specified that every effort would be made to maintain participant confidentiality, including through deidentifying private information. Participants were compensated US $25 for each study visit and up to US $150 for completion of all study visits.

Results

Overview

The trial randomized 351 pregnant people, and 348 had complete baseline data (Tables S4 and S5 in Multimedia Appendix 1). The 348 individuals were reduced to 258 individuals for analysis, as 85 individuals did not complete the final visit and 5 did not have late pregnancy household chaos scores (Figure 1). There was no difference in early pregnancy household chaos score between those included (mean 25.9, SD 7.4; n=258) and those not included due to missing data (mean 26.6, SD 9.2; n=60).

Figure 1. CONSORT diagram for illustrating participant progression through the trial.



As shown in Table 1, there was a comparable distribution across treatment groups. Participants were primarily non-Hispanic Black (139/258, 53.9%), married or living with a significant other (147/258, 57.0%), and nulliparous (112/258, 43.4%). There was a significant difference in the parity distribution between treatment groups (*P*=.02), but no other differences in maternal characteristics or health outcomes (All *P*>.05).

Table 1. Demographic characteristics of included pregnant people (N=258).

Maternal characteristics	All (N=258)	Intervention (n=135)	Usual care (n=123)	*P* value[a]
Maternal age (years), mean (SD)	27.5 (6.0)	27.5 (6.0)	27.5 (6.0)	.99
Weight (kg), mean (SD)	74.8 (16.9)	74.2 (16.4)	75.3 (16.5)	.58
BMI (kg/m^2), mean (SD)	32.2 (5.0)	31.8 (5.0)	32.6 (5.0)	.18
Normal (18.5-24.9), n (%)	89 (34.5)	47 (34.8)	42 (34.1)	.99
Overweight (25-29.9), n (%)	76 (29.5)	36 (26.7)	40 (29.6)	—[b]
Obese (30-40), n (%)	93 (36.0)	48 (35.6)	45 (36.6)	—
Race and ethnicity, n (%)				.35
Hispanic	18 (7.0)	7 (5.2)	11 (8.9)	
Non-Hispanic White	85 (32.9)	45 (33.3)	40 (32.5)	
Non-Hispanic Black	139 (53.9)	77 (57.0)	62 (50.4)	
Mixed or other	16 (6.2)	6 (4.4)	10 (8.1)	
Marital status, n (%)				.30
Married/living with significant other	147 (57.0)	81 (60.0)	66 (53.7)	
Not married	111 (43.0)	54 (40.0)	57 (46.3)	
Education level, n (%)				.91
Postgraduate work	9 (3.5)	4 (3.0)	5 (4.1)	
College degree	31 (12.0)	17 (12.6)	14 (11.4)	
1-3 years of college or technical school	128 (49.6)	64 (47.4)	64 (52.0)	
High school diploma or equivalent	74 (28.7)	41 (30.4)	33 (26.8)	
Some high school	16 (6.2)	9 (6.7)	7 (5.7)	
Parity, n (%)				.02[c]
0	112 (43.4)	63 (46.7)	49 (39.8)	
1	69 (26.7)	27 (20.0)	42 (34.1)	
2	34 (13.2)	24 (17.8)	10 (8.1)	
3+	43 (16.7)	21 (15.6)	22 (17.9)	
Enrollment time, mean (SD)				.70
Pre–COVID-19 (2019-March 2020)	15 (5.8)	8 (5.9)	7 (5.7)	
COVID-19 (March 2020-March 2021)	41 (15.9)	19 (14.1)	22 (17.9)	
Post–COVID-19 (April 2021-2023)	202 (78.3)	108 (80.0)	94 (76.4)	
Gestational diabetes, n (%)	19 (7.4)	9 (6.7)	10 (8.1)	.68
Baseline stress, mean (SD)	9.7 (8.7)	9.7 (8.9)	9.3 (8.2)	.69

[a]Comparisons between groups were conducted using chi-square (categorical variables) or linear regression (continuous variables).

[b]Not applicable.

[c]$P<.05$.

Aim 1: Household Chaos Across Pregnancy

Across both treatment groups combined, most individuals were classified as having low (score <25, n=124, 48.1%) or moderate or low (score 25-30, n=73, 28.3%) household chaos in early pregnancy (Table 2). There were minimal changes in household chaos scores from early to late pregnancy, although scores and categories tended to be higher in late pregnancy (Table 2). Similarly, there were no significant changes in depression or stress scores, although there was a small improvement in anxiety among the full sample (Table S6 in Multimedia Appendix 1).

Table 2. Changes in household chaos across gestation (n=258).

	Early pregnancy	Late pregnancy	Change	*P* value[a]
Total score, mean (SD)	25.9 (7.4)	26.4 (8.5)	0.49 (8.27)	.34
Category, n (%)			—[b]	.21
Low (<25)	124 (48.1)	128 (49.6)		
Moderate-low (25-30)	73 (28.3)	70 (27.1)		
Moderate-high (31-35)	36 (14.0)	22 (8.5)		
High (36-60)	25 (9.7)	37 (14.3)		
Household chaos improved or no change (n=140)				
Total score, mean (SD)	27.9 (7.6)	23.4 (6.2)	–4.4 (5.0)	.01[c]
Category, n (%)			—	.01[c]
Low (<25)	49 (35.0)	86 (61.4)		
Moderate-low (25-30)	48 (34.3)	40 (28.6)		
Moderate-high (31-35)	22 (15.7)	6 (4.3)		
High (36-60)	21 (15.0)	8 (5.7)		
Household chaos worsened (n=118)				
Total score, mean (SD)	23.6 (6.3)	30.0 (9.4)	6.3 (7.4)	.01[c]
Category, n (%)			—	.01[c]
Low (<25)	75 (63.6)	42 (35.6)		
Moderate-low (25-30)	25 (21.2)	30 (25.4)		
Moderate-high (31-35)	14 (11.9)	16 (13.6)		
High (36-60)	4 (3.4)	29 (24.6)		

[a]Comparisons between groups were conducted using the McNemar test (categorical variables) or paired *t* test (continuous variables).

[b]Not applicable.

[c]$P<.05$.

Participants were categorized into improvement groups: either household chaos improving or no change, or worsening (Table 2). Approximately half of the 258 participants (n=140, 54.3%) had their household chaos improve (n=109, 42.2%) or had no change (n=31, 12.0%), and the others had their household chaos worsen (n=118, 45.7%). Both improvement groups had a mean change of around 5 points, either an increase or decrease, which resulted in shifts across the 5-point categories ($P<.05$). Using two SDs of this population mean (14.60 units) [30], we found only 6.2% (n=16) participants reported changes more than two SDs, with 3.1% (n=8) of the sample reporting a clinically significant decrease in household chaos scores and 3.1% (n=8) of the reporting a clinically significant increase in household chaos scores. Participants with improved or no change in household chaos scores were mostly categorized in the moderate-to-low chaos category in early pregnancy, but were categorized as low chaos by late pregnancy scores (n=158, 61.2%). They also began with a higher (or worse) early pregnancy household chaos score (mean 27.9, SD 7.6) compared to those in the worsening group (mean 23.6, SD 6.3). The worsening group was the converse, whereby the majority began in the low chaos category and were categorized as moderate-to-low chaos category in late pregnancy. Only early pregnancy chaos was associated with the improvement category. There were no other significant relationships between early pregnancy covariates and household chaos improvement category or change (all $P>.05$).

Within BMI categories independent of treatment assignment, there were no significant changes in household chaos scores across time (all $P>.05$; Table S6 in Multimedia Appendix 1). In individuals who were overweight at enrollment (n=76), early pregnancy weight was positively correlated with early pregnancy household chaos scores ($r=0.32$; $P=.004$) and negatively correlated with household chaos change ($r=–0.22$; $P=.05$). Therefore, people with a higher weight within the overweight category reported a higher household chaos score in early pregnancy but lowered their chaos over time. There were no other significant associations between early pregnancy scores, household chaos change, and weight outcomes in BMI categories.

When examining within-treatment groups and not BMI categories, there were no significant changes in household chaos across time (all $P>.05$; Table S6 in Multimedia Appendix 1). In the intervention group (n=135), GWG was negatively associated with late pregnancy household chaos ($r=–0.17$; $P=.03$). Therefore, intervention group participants who gained less weight had a higher late pregnancy household chaos score.

Similarly, in the intervention group, early pregnancy household chaos score was negatively associated with GWG (r=–0.13; P=.12), but this was not statistically significant. There were no other significant effects between household chaos scores at either time point or the change thereof and weight outcomes in treatment groups.

Aim 2: Mediation Model of Intervention, Household Chaos, and Gestational Weight Gain

In the adjusted models, household chaos change did not differ by treatment (P=.74). Both early pregnancy household chaos (β=–0.53, SE 0.06; P=.001) and DASS-21–assessed stress (β=0.15, SE 0.05; P=.004) were associated with household chaos change in these models. Further, household chaos change was not associated with GWG (P=.96). Higher parity (β=–1.06, SE 0.40; P=.009), BMI category (β=–2.49, SE 0.39; P=.001), and gestational diabetes status (β=–2.66, SE 1.27; P=.03) was negatively associated with GWG (all P<.05). A mediation model was not pursued, as there was no association between the predictor and mediator or mediator and outcome.

Discussion

Principal Findings

The purpose of this paper was to examine household chaos across pregnancy in a sample of pregnant people with lower SES and to examine whether additional household chaos negatively impacted the effect of a behavioral lifestyle intervention on GWG. In this study, people began pregnancy with generally low or moderate-to-low household chaos and with differing, although minimal, progression across gestation. The individual-focused eHealth behavioral intervention did not change household chaos score, nor was the amount of household chaos or its change related to GWG.

To our knowledge, there are limited reports of household chaos during pregnancy [13]. The observed mean early and late pregnancy household chaos scores were comparable to lower-income mothers of 12-month-old infants (mean 25.1, SD 6.7) [31], but lower than parents of young children (score 31) [32]. These lower household chaos scores may be attributed to around half of the current sample (112/258, 43.4%) being nulliparous. Accordingly, a minority of the current sample was classified as "low" chaos at either point in gestation compared to other samples using similar chaos categories: parents of young children (27%) [23], and mothers of young children during the COVID-19 pandemic (20%) [33]. Moreover, fewer pregnant people in this sample were categorized as high chaos in early pregnancy (~11%) compared to these past investigations of primarily high-income parents of young children (range 24%-27%) [23,33]. This finding is surprising, given a mixed methods study in this same state found that the COVID-19 pandemic and hurricanes negatively impacted pregnant people's home environment and their mental health, especially pregnant people with low SES [34]. Two immediate explanations may be presented. First, only pregnant people who live in homes with low chaos may enroll in such intensive lifestyle intervention due to additional support at home or time available; therefore, these findings may be a result of selection bias and lower early pregnancy chaos serving as a facilitator to participation. Further, people in high-chaos homes may have delivered early; indeed, distress during pregnancy is a known risk factor for preterm birth [35]. Even so, there was no difference in the early pregnancy household chaos score between those who were included and those not included in the analysis. The second consideration is that household chaos scores significantly increase after the birth of the child, which may align with the notable prevalence and impacts of postpartum depression [36]. This increase may be sustained as the child ages, as higher parity was associated with a higher early pregnancy household chaos score in this sample.

We did not find support for our first hypothesis that household chaos scores would increase across gestation; we did not observe changes in household chaos across gestation when explored continuously, and there was an even split of improvement and worsening of household chaos. The first consideration is that the finding could reflect regression to the mean, whereby the data may move closer to the mean after an extreme value. Still, these results echo minimal changes across 12 months (1 point) in a longitudinal investigation in infants of lower-income homes [31]. Using a two SD cutoff of the current sample, we found very few made clinically meaningful changes. However, these results assume that this sample is a functional or normal population [30]; yet, no such normative data exist in pregnancy. Therefore, it is unclear if the 5-point changes in both improvement and worsening groups could also be clinically meaningful, as other longitudinal investigations have primarily examined similar median mean splits [37] or used differing scoring for the current questionnaire [16,38,39]. Changes between improvement groups indicated that those with moderate-to-low chaos improved their environment, while individuals in the low household chaos category did not sustain their routine environment. After months of additional anxiety, parents may make changes to cope and reduce anxiety before the arrival of the baby. These changes mirror another longitudinal investigation whereby stress at 6 months post partum led to improvements in sleep duration at 8 months post partum [40]. Another consideration is that the explored covariates (eg, age at enrollment and parity) did not change as pregnancy progressed. Investigating time-varying covariates, such as income and employment changes, may improve upon these practices.

We did not find support for our second hypothesis, that household chaos change would mediate the relationship between the intervention and GWG. Rather, across household chaos categories and despite any household chaos change, pregnant people were able to gain less weight in the intervention. These null results align with another obesity-focused intervention in 77 families of young children (ages 18 months-5 years) that showed minimal change (~1 point) after a 6-month intervention and 1-year follow-up [32], but contrast a routine-focused intervention in 54 parents of young children (ages 2-4 years) that decreased household chaos within 12 weeks [16]. Others have hypothesized that the relationship between stress and GWG is primarily explained by demographic factors, like low SES and lower education [41,42]. Our analysis did find that early pregnancy stress was a key covariate between the intervention group and household chaos change. This may suggest that early pregnancy stress is a mediator of household chaos change. It is possible that early pregnancy stress may negatively impact household chaos or, in turn, lead to an increase in household chaos. This current sample's common factor of low SES may build upon other higher-income GWG interventions, but may also preclude associations between stress and GWG. Examination of specific health behaviors across the intervention may further elucidate the mechanism of household chaos and gaining an appropriate amount of weight in pregnancy [43], as household chaos is directly linked to maternal sleep in mothers of young children [15]. In this intervention, it is possible that participants adopted a healthier lifestyle, as indicated by gaining less GWG, which may have positive implications for their future parenting practices and reduce the impact of household chaos on child development.

Strengths and Limitations

Strengths of this study include the innovative examination of household chaos across pregnancy, an evidence-based pragmatic eHealth lifestyles intervention that occurred within a vulnerable population, and the concurrent measurement of stress concepts. Limitations include the lack of assessment of other covariates, missing household chaos data for some individuals in late pregnancy, self-report of chaos measures, difficulty identifying active intervention components, and potential lack of generalizability to other pregnant populations. First, additional information on household and economic stability [7], including food insecurity [44], employment security, and family composition (ie, number of children), may help identify further contributors to household chaos in this population. Nevertheless, the current assessment of household chaos did capture parity, and the household chaos score was correlated with other measures of mental health in this study (eg, depression, anxiety, and stress), validating its assessment of poor mental health. Second, identifying active intervention ingredients was difficult as the intervention addressed critical concepts (eg, stress reduction, routine, and the home environment) in multiple intervention sessions. This multicomponent approach allowed for significant intervention effects on GWG [18], but it limited our ability to thoroughly examine intervention components on maternal mental health. Moreover, this analysis was conducted by group assignment, and further investigation into how higher or lower amounts of household chaos may impact engagement in the intervention is warranted. Third, these results are confined to a self-reported questionnaire, which provides the opportunity for social desirability bias to arise, potentially resulting in the reporting of lower household chaos scores. Given the intervention's eHealth modality and the ensuing COVID-19 pandemic, an existing validated questionnaire was conducted using a paper form rather than delivered during an in-home or psychologist visit. Finally, the current investigation is confined to pregnant people without current unstable mental health conditions who were enrolled in a supplemental nutrition program in a southern US state. These results may not apply to pregnant people with current mental health disorders (eg, prenatal depression), not participating in WIC, or living in other US or global regions.

Future Directions

The current results suggest 4 major areas for future research, policy, and practice. First, household chaos assessment across pregnancy and postpartum may help identify critical time points for intervention and such normative data for clinically meaningful cutoffs. Second, evaluating the effect of in-person intervention on household chaos may improve upon the current study's eHealth-based intervention. Though in-person interventions may result in lower compliance [12], they may foster higher social support and changes at home. Third, creating a household chaos program for pregnancy may still be warranted for high-chaos households or to prepare for the higher chaos when the baby arrives. This potential program may include focusing on household routines, reducing household screen time, and promoting family cohesion for parent and family stress reduction [16]. Moreover, reducing any household chaos in the postpartum period may help support maternal mental and physical health. Fourth, creating a multilevel program to improve economic and housing stability may reduce household stress and improve upon this individual-focused intervention. Even so, it may be seen as a benefit that the current intervention did not result in more household chaos in this population.

Conclusion

In this sample of pregnant people with lower SES, a low level of household chaos was sustained across gestation. Accordingly, this intervention was related to lower GWG, but not their household chaos scores or change across 6 months. Beneficial changes in lifestyle behaviors have positive implications for adequate weight gain, reducing stress, and future positive parenting practices. Routine-focused and multilevel interventions may improve upon these findings to support an organized home for future parent and child health.

Acknowledgments

The authors are indebted to the support from the Louisiana Department of Health; the Louisiana Women's, Infants, and Children's program staff; and the participants.

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding

This research was supported by funding from the National Institute of Nursing Research (5R01NR017644), the Louisiana/Pennington Nutrition and Obesity Research Center of the National Institutes of Diabetes, Digestive and Kidney Diseases (P30DK072476), and the Louisiana Clinical and Translational Sciences Center (U54 GM104940). CLK was supported by the National Institute of General Medical Sciences (P20GM144269). EKW is currently supported by the National Heart, Lung, and Blood Institute (F32HL176217). EWF was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (F32HD108022). HEC was supported by the National Institute of Diabetes and Digestive and Kidney Diseases (T32DK064584). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Authors' Contributions

CLK developed the initial research question, carried out data analyses, drafted the initial manuscript, and reviewed and revised the manuscript. KF helped develop the research question, carried out data analyses, and reviewed and revised the manuscript. CLK wrote the initial manuscript draft with contributions from EWF, HEC, MK, EKW, and JWA. ADA, RB, and DSH provided study oversight and helped interpret results. LR conceptualized and designed the *Smart Moms in WIC* study, designed the data collection instruments, helped develop the research question and analysis plan, and critically reviewed and revised the manuscript. All authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
SmartMoms SmartTips by week, covariates of baseline household chaos, early pregnancy mental health correlations, participant demographics, and household chaos changes by BMI and treatment group.
[DOCX File , 37 KB - mental_v13i1e74146_app1.docx]

Multimedia Appendix 2
CONSORT e-Health Checklist (V 1.6.1).
[PDF File (Adobe PDF File), 15907 KB - mental_v13i1e74146_app2.pdf]

References

1. Barker DJP. The origins of the developmental origins theory. J Intern Med 2007;261(5):412-417 [FREE Full text] [doi: 10.1111/j.1365-2796.2007.01809.x] [Medline: 17444880]
2. McEwen B. Stress, adaptation, and disease. Allostasis and allostatic load. Ann N Y Acad Sci 1998;840:33-44. [doi: 10.1111/j.1749-6632.1998.tb09546.x] [Medline: 9629234]
3. Traylor CS, Johnson JD, Kimmel MC, Manuck TA. Effects of psychological stress on adverse pregnancy outcomes and nonpharmacologic approaches for reduction: an expert review. Am J Obstet Gynecol MFM 2020;2(4):100229 [FREE Full text] [doi: 10.1016/j.ajogmf.2020.100229] [Medline: 32995736]
4. Van den Bergh BR, van den Heuvel MI, Lahti M, et al. Prenatal developmental origins of behavior and mental health: the influence of maternal stress in pregnancy. Neurosci Biobehav Rev 2020;117:26-64. [doi: 10.1016/j.neubiorev.2017.07.003] [Medline: 28757456]
5. Miller AL, Lumeng JC. Pathways of association from stress to obesity in early childhood. Obesity (Silver Spring) 2018;26(7):1117-1124 [FREE Full text] [doi: 10.1002/oby.22155] [Medline: 29656595]
6. Kraft P, Kraft B. Explaining socioeconomic disparities in health behaviours: a review of biopsychological pathways involving stress and inflammation. Neurosci Biobehav Rev 2021;127:689-708 [FREE Full text] [doi: 10.1016/j.neubiorev.2021.05.019] [Medline: 34048858]
7. Brown ED, Anderson KE, Garnett ML, Hill EM. Economic instability and household chaos relate to cortisol for children in poverty. J Fam Psychol 2019;33(6):629-639. [doi: 10.1037/fam0000545] [Medline: 31169392]
8. Zimmermann M, Julce C, Sarkar P, et al. Can psychological interventions prevent or reduce risk for perinatal anxiety disorders? A systematic review and meta-analysis. Gen Hosp Psychiatry 2023;84:203-214. [doi: 10.1016/j.genhosppsych.2023.08.005] [Medline: 37619299]
9. Lau Y, Chew HSJ, Ang WHD, et al. Effects of digital health interventions on the psychological outcomes of perinatal women: umbrella review of systematic reviews and meta-analyses. Health Psychol Rev 2024;18(2):229-254 [FREE Full text] [doi: 10.1080/17437199.2023.2185654] [Medline: 36919443]
10. Sakamoto JL, Carandang RR, Kharel M, et al. Effects of mHealth on the psychosocial health of pregnant women and mothers: a systematic review. BMJ Open 2022;12(2):e056807 [FREE Full text] [doi: 10.1136/bmjopen-2021-056807] [Medline: 35168981]

11. Bogaerts AFL, Devlieger R, Nuyts E, Witters I, Gyselaers W, Van den Bergh BRH. Effects of lifestyle intervention in obese pregnant women on gestational weight gain and mental health: a randomized controlled trial. Int J Obes (Lond) 2013;37(6):814-821. [doi: 10.1038/ijo.2012.162] [Medline: 23032404]
12. Garnæs KK, Helvik AS, Stafne SN, et al. Effects of supervised exercise training during pregnancy on psychological well-being among overweight and obese women: secondary analyses of the ETIP-trial, a randomised controlled trial. BMJ Open 2019;9(11):e028252 [FREE Full text] [doi: 10.1136/bmjopen-2018-028252] [Medline: 31753866]
13. Marsh S, Dobson R, Maddison R. The relationship between household chaos and child, parent, and family outcomes: a systematic scoping review. BMC Public Health 2020;20(1):513 [FREE Full text] [doi: 10.1186/s12889-020-08587-8] [Medline: 32316937]
14. Matheny AP, Wachs TD, Ludwig JL, Phillips K. Bringing order out of chaos: Psychometric characteristics of the confusion, hubbub, and order scale. Journal of Applied Developmental Psychology 1995;16(3):429-444. [doi: 10.1016/0193-3973(95)90028-4]
15. Kracht CL, Katzmarzyk PT, Staiano AE. Household chaos, maternal stress, and maternal health behaviors in the United States during the COVID-19 outbreak. Womens Health (Lond) 2021;17:17455065211010655 [FREE Full text] [doi: 10.1177/17455065211010655] [Medline: 33886392]
16. Marsh S, Taylor R, Galland B, Gerritsen S, Parag V, Maddison R. Results of the 3 Pillars Study (3PS), a relationship-based programme targeting parent-child interactions, healthy lifestyle behaviours, and the home environment in parents of preschool-aged children: A pilot randomised controlled trial. PLoS One 2020;15(9):e0238977 [FREE Full text] [doi: 10.1371/journal.pone.0238977] [Medline: 32941530]
17. Dhaliwal SK, Dabelea D, Lee-Winn AE, Crume T, Wilkening G, Perng W. Maternal psychosocial stress during pregnancy and offspring neurobehavioral outcomes during early childhood in the healthy start study. Ann Epidemiol 2023;86:16-24.e3. [doi: 10.1016/j.annepidem.2023.06.001] [Medline: 37321280]
18. Flanagan EW, Falkenhain K, Beyl R, Altazan AD, Richard SA, Cabre HE. Gestational weight gain management in underserved mothers - a state-wide randomized controlled trial in louisiana wic. medRxiv 2025. [doi: 10.1101/2025.01.29.25321347]
19. Flanagan EW, Altazan AD, Comardelle NR, et al. The design of a randomized clinical trial to evaluate a pragmatic and scalable ehealth intervention for the management of gestational weight gain in low-income women: protocol for the smartmoms in WIC Trial. JMIR Res Protoc 2020;9(9):e18211 [FREE Full text] [doi: 10.2196/18211] [Medline: 32909954]
20. Institute of Medicine (US) and National Research Council (US) Committee to Reexamine IOM Pregnancy Weight Guidelines. Weight gain during pregnancy: reexamining the guidelines. In: Rasmussen KM, Yaktine AL, editors. The National Academies Collection: Reports funded by National Institutes of Health. Washington (DC): National Academies Press; 2009.
21. Institute of Medicine (U.S.). Committee to reexamine IOM pregnancy weight guidelines. In: Rasmussen KM, Yaktine AL, editors. Weight Gain During Pregnancy : Reexamining the Guidelines. Washington, DC: National Academies Press; 2009.
22. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
23. Emond JA, Tantum LK, Gilbert-Diamond D, Kim SJ, Lansigan RK, Neelon SB. Household chaos and screen media use among preschool-aged children: a cross-sectional study. BMC Public Health 2018;18(1):1210 [FREE Full text] [doi: 10.1186/s12889-018-6113-2] [Medline: 30373557]
24. Crawford JR, Henry JD. The depression anxiety stress scales (DASS): normative data and latent structure in a large non-clinical sample. Br J Clin Psychol 2003;42(Pt 2):111-131. [doi: 10.1348/014466503321903544] [Medline: 12828802]
25. Migueles JH, Rowlands AV, Huber F, Sabia S, van HVT. GGIR: A research community?Driven open source r package for generating physical activity and sleep outcomes from multi-day raw accelerometer data. Journal for the Measurement of Physical Behaviour 2019;2(3):188-196. [doi: 10.1123/jmpb.2018-0063]
26. van Hees VT, Sabia S, Anderson KN, et al. A novel, open access method to assess sleep duration using a wrist-worn accelerometer. PLoS One 2015;10(11):e0142533 [FREE Full text] [doi: 10.1371/journal.pone.0142533] [Medline: 26569414]
27. van Hees VT, Sabia S, Jones SE, et al. Estimating sleep parameters using an accelerometer without sleep diary. Sci Rep 2018;8(1):12975 [FREE Full text] [doi: 10.1038/s41598-018-31266-z] [Medline: 30154500]
28. Johnson AD, Martin A, Partika A, et al. Chaos during the COVID-19 outbreak: predictors of household chaos among low-income families during a pandemic. Fam Relat 2022;71(1):18-28 [FREE Full text] [doi: 10.1111/fare.12597] [Medline: 34898781]
29. Hayes AF. Introduction to Mediation, Moderation, and Conditional Process Analysis : A Regression-Based Approach. New York, NY: Guilford Press; 2018:692.
30. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59(1):12-19. [doi: 10.1037//0022-006x.59.1.12] [Medline: 2002127]
31. Khatiwada A, Shoaibi A, Neelon B, Emond JA, Benjamin-Neelon SE. Household chaos during infancy and infant weight status at 12 months. Pediatr Obes 2018;13(10):607-613 [FREE Full text] [doi: 10.1111/ijpo.12395] [Medline: 30019385]

32. Hruska V, Darlington G, Haines J, Ma DWL. Parent stress as a consideration in childhood obesity prevention: results from the guelph family health study, a pilot randomized controlled trial. Nutrients 2020;12(6):1835 [FREE Full text] [doi: 10.3390/nu12061835] [Medline: 32575660]
33. Kracht CL, Katzmarzyk PT, Staiano AE. Household chaos, family routines, and young child movement behaviors in the U.S. during the COVID-19 outbreak: a cross-sectional study. BMC Public Health 2021;21(1):860 [FREE Full text] [doi: 10.1186/s12889-021-10909-3] [Medline: 33947357]
34. Kracht CL, Goynes KO, Dickey M, et al. The role of government assistance, housing, and employment on postpartum maternal health across income and race: a mixed methods study. BMC Public Health 2024;24(1):3244 [FREE Full text] [doi: 10.1186/s12889-024-20745-w] [Medline: 39574054]
35. Staneva A, Bogossian F, Pritchard M, Wittkowski A. The effects of maternal depression, anxiety, and perceived stress during pregnancy on preterm birth: a systematic review. Women Birth 2015;28(3):179-193. [doi: 10.1016/j.wombi.2015.02.003] [Medline: 25765470]
36. Slomian J, Honvo G, Emonts P, Reginster J, Bruyère O. Consequences of maternal postpartum depression: a systematic review of maternal and infant outcomes. Womens Health (Lond) 2019;15:1745506519844044 [FREE Full text] [doi: 10.1177/1745506519844044] [Medline: 31035856]
37. Whitesell CJ, Crosby B, Anders TF, Teti DM. Household chaos and family sleep during infants' first year. J Fam Psychol 2018;32(5):622-631 [FREE Full text] [doi: 10.1037/fam0000422] [Medline: 29781634]
38. Asta K, Miller A, Retzloff L, Rosenblum K, Kaciroti N, Lumeng J. Eating in the absence of hunger and weight gain in low-income toddlers. Pediatrics 2016;137(5) [FREE Full text] [doi: 10.1542/peds.2015-3786] [Medline: 27244808]
39. McQuillan ME, Bates JE, Staples AD, Deater-Deckard K. A 1-year longitudinal study of the stress, sleep, and parenting of mothers of toddlers. Sleep Health 2022;8(1):47-53 [FREE Full text] [doi: 10.1016/j.sleh.2021.08.006] [Medline: 34620578]
40. Kracht CL, Blanchard CM, Symons Downs D, Beauchamp MR, Rhodes RE. New parents' sleep, movement, health, and well-being across the postpartum period. Behav Sleep Med 2024;22(5):636-649. [doi: 10.1080/15402002.2024.2339815] [Medline: 38592976]
41. Vehmeijer FOL, Balkaran SR, Santos S, et al. Psychological distress and weight gain in pregnancy: a population-based study. Int J Behav Med 2020;27(1):30-38 [FREE Full text] [doi: 10.1007/s12529-019-09832-0] [Medline: 31853868]
42. O'Brien E, Alberdi G, McAuliffe F. The influence of socioeconomic status on gestational weight gain: a systematic review. J Public Health (Oxf) 2018;40(1):41-55. [doi: 10.1093/pubmed/fdx038] [Medline: 28398550]
43. Geiker NRW, Astrup A, Hjorth MF, Sjödin A, Pijls L, Markus CR. Does stress influence sleep patterns, food intake, weight gain, abdominal obesity and weight loss interventions and vice versa? Obes Rev 2018;19(1):81-97. [doi: 10.1111/obr.12603] [Medline: 28849612]
44. Nguyen G, Bell Z, Andreae G, et al. Food insecurity during pregnancy in high-income countries, and maternal weight and diet: a systematic review and meta-analysis. Obes Rev 2024;25(7):e13753. [doi: 10.1111/obr.13753] [Medline: 38693587]

Abbreviations

DASS-21: 21-item Depression, Anxiety, and Stress Scale
GWG: gestational weight gain
SES: socioeconomic status
WIC: Supplemental Nutrition Program for Women, Infants, and Children

Edited by J Torous; submitted 18.Mar.2025; peer-reviewed by J Emond, M Waring; comments to author 29.Jul.2025; revised version received 08.Sep.2025; accepted 02.Oct.2025; published 08.Jan.2026.

Please cite as:
Kracht CL, Falkenhain K, Flanagan EW, Altazan AD, Cabre HE, Kebbe M, Woolf EK, Beyl R, Hsia DS, Apolzan JW, Redman L
Effect of a Pragmatic eHealth Behavioral Gestational Weight Gain Intervention on Household Chaos in Pregnant People of Lower Socioeconomic Status: Randomized Controlled Trial
JMIR Ment Health 2026;13:e74146
URL: https://mental.jmir.org/2026/1/e74146
doi:10.2196/74146
PMID:

Prediction of 12-Week Remission in Patients With Depressive Disorder Using Reasoning-Based Large Language Models: Model Development and Validation Study

Jin-Hyun Park[1*], MSc; Hee-Ju Kang[2*], MD, PhD; Ji Hyeon Jeon[2], MD; Sung-Gil Kang[2], MD; Ju-Wan Kim[2], MSc, MD; Jae-Min Kim[2], MD, PhD; Hwamin Lee[1], PhD

[1]Department of Biomedical Informatics, Korea University College of Medicine, 161, Jeongneung-ro, Seongbuk-gu, Seoul, Republic of Korea

[2]Department of Psychiatry, Chonnam National University Medical School, Gwangju, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Hwamin Lee, PhD
Department of Biomedical Informatics, Korea University College of Medicine, 161, Jeongneung-ro, Seongbuk-gu, Seoul, Republic of Korea

Abstract

Background: Depressive disorder affects over 300 million people globally, with only 30% to 40% of patients achieving remission with initial antidepressant monotherapy. This low response rate highlights the critical need for digital mental health tools that can identify treatment response early in the clinical pathway.

Objective: This study aimed to evaluate whether reasoning-based large language models (LLMs) could accurately predict 12-week remission in patients with depressive disorder undergoing antidepressant monotherapy and to assess the clinical validity and interpretability of model-generated rationales for integration into digital mental health workflows.

Methods: We analyzed data from 390 patients in the MAKE Biomarker discovery study who were undergoing first-step antidepressant monotherapy with 12 different medications, including escitalopram, paroxetine, sertraline, duloxetine, venlafaxine, desvenlafaxine, milnacipran, mirtazapine, bupropion, vortioxetine, tianeptine, and trazodone, after excluding those with uncommon medications (n=9) or missing biomarker data (n=32). Three LLMs (ChatGPT o1, o3-mini, and Claude 3.7 Sonnet) were tested using advanced prompting strategies, including zero-shot chain-of-thought, atom-of-thoughts, and our novel referencing of deep research prompt. Model performance was evaluated using balanced accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. Three psychiatrists independently assessed model outputs for clinical validity using 5-point Likert scales across multiple dimensions.

Results: Claude 3.7 Sonnet with 32,000 reasoning tokens using the referencing of deep research prompt achieved the highest performance (balanced accuracy=0.6697, sensitivity=0.7183, and specificity=0.6210). Medication-specific analysis revealed negative predictive values of 0.75 or higher across major antidepressants, indicating particular utility in identifying likely nonresponders. Clinical evaluation by psychiatrists showed favorable mean ratings for correctness (4.3, SD 0.7), consistency (4.2, SD 0.8), specificity (4.2, SD 0.7), helpfulness (4.2, SD 1.0), and human likeness (3.6, SD 1.7) on 5-point scales.

Conclusions: These findings demonstrate that reasoning-based LLMs, particularly when enhanced with research-informed prompting, show promise for predicting antidepressant response and could serve as interpretable adjunctive tools in depressive disorder treatment planning, although prospective validation in real-world clinical settings remains essential.

(JMIR Ment Health 2026;13:e83352) doi:10.2196/83352

KEYWORDS

artificial intelligence; clinical support systems; depressive disorder; large language models; natural language processing; prognosis; treatment outcome

Introduction

Depressive disorder is one of the most prevalent and debilitating psychiatric conditions worldwide, ranking as a primary contributor to global disability and significantly influencing the overall disease burden associated with mental disorders [1]. Given the substantial burden imposed by depressive disorder, optimizing strategies for early diagnosis, effective treatment, and personalized intervention remains a critical public health priority. Despite the critical need for effective intervention, the primary treatment objective of achieving remission, defined as near-complete symptom resolution, remains challenging, with initial antidepressant monotherapy resulting in remission rates of only 30% to 40% within 12 weeks [2,3]. This limited success often necessitates multiple treatment trials, consequently prolonging suffering, increasing health care use and suicide risk, elevating dropout rates [4], and ultimately exacerbating patient distress while significantly amplifying treatment nonadherence [5].

Consequently, the early identification of patients who will not achieve remission with a particular monotherapy regimen has become a critical topic in both research and clinical practice. Early identification of patients who are less likely to respond to standard first-line treatments would allow clinicians to tailor interventions more efficiently and reduce the time lost during ineffective treatments [6]. Recent studies have explored the use of machine learning (ML) models to predict remission in patients with depressive disorder. However, these investigations have encountered limitations, resulting from study design, which may not reflect real-world clinical practice, including limited diversity in the antidepressants administered and challenges in clinically interpreting the predictions generated by ML models [7-10].

In recent developments, large language models (LLMs) have emerged as promising instruments for various psychiatric applications, encompassing diagnostic assessment, risk stratification, and clinical decision support [11-13]. Furthermore, LLMs that enhance chain-of-thought reasoning, such as OpenAI’s ChatGPT o1 [14], ChatGPT o3-mini [15], and Anthropic’s Claude 3.7 Sonnet [16], have been developed and applied within the medical field to improve diagnostic reasoning. These reasoning-enhanced LLMs have demonstrated potential across various medical specialties, yet their application to predicting antidepressant treatment outcomes remains unexplored [17-21].

Therefore, in this study, we aimed to evaluate whether reasoning-enhanced LLMs could accurately predict 12-week remission among patients with depressive disorder undergoing monotherapy with 1 of 12 different antidepressants, including selective serotonin reuptake inhibitors (SSRIs), serotonin and norepinephrine reuptake inhibitors (SNRIs), or other antidepressants. We also investigated the underlying clinical rationale of these predictions and explored the feasibility of proposing alternative treatment strategies when remission was deemed unlikely.

Methods

Participants and Data Preprocessing

The dataset for this study was obtained from the MAKE Biomarker Discovery for Enhancing Antidepressant Treatment Effect and Response (MAKE BETTER) study [22]. Patients with depressive disorders were consecutively recruited from March 2012 to April 2017 at the outpatient psychiatry department of Chonnam National University Hospital. From the initial cohort, 431 patients who received first-step monotherapy were identified. After excluding 9 patients prescribed “other” medications and 32 lacking blood biomarker data, a total of 390 patients were included in the final analysis.

Variables assessed included demographic characteristics, personal and familial psychiatric histories, comorbidities, responses to the 9-item Mini-International Neuropsychiatric Interview [23], adverse childhood experiences before the age of 16 years (physical, psychological, and sexual abuse), depression subtypes (including melancholic, atypical, and psychotic), and prescribed antidepressants and dosage. Suicidality was assessed using a structured interview comprising 4 standardized questions addressing suicidal thoughts and intent (eg, “Have you ever felt that life is not worth living?”). The presence of suicidal ideation determined from these structured questions was subsequently reflected in the Brief Psychiatric Rating Scale [24] suicidality item rating. For analysis, only the binary presence or absence of suicidal ideation was used, not the raw Brief Psychiatric Rating Scale score. Additional variables included the Hamilton Depression Rating Scale (HAM-D) [25] score, health-related quality of life (EQ-5D) [26], functional impairment (Sheehan Disability Scale) [27], perceived stress (Perceived Stress Scale) [28], resilience (Conner-Davidson Resilience Scale) [29], perceived social support (Multidimensional Scale of Perceived Social Support) [30], blood biomarkers at baseline, and early treatment response at 2 weeks (≥20% reduction in HAM-D scores). For female participants, fertility and depression-related factors were evaluated, including age at menarche or menopause, hormonal therapy use, and presence of peri- or postpartum or postmenopausal depression. Further details on eligibility, pharmacotherapy, clinical assessments, and biomarker procedures are provided in Multimedia Appendix 1. The primary outcome was 12-week remission, defined as an HAM-D score ≤7 sustained through the 12-week assessment point. All analyzed participants were adults, consistent with the validated use of psychiatric assessment tools and pharmacotherapy in adult outpatient clinical practice.

Numeric coded data were transformed into structured, narrative-style reports in natural language to enhance interpretability by the LLMs, and the comprehensive structure of patient information is depicted in Textbox 1.

Textbox 1. Structured representation of patient information used for input to the large language models (LLMs). This figure illustrates the structured format of patient information for individuals with major depressive disorder as prepared for LLM input. Each patient's clinical data were inserted into the (patient information) section of the experimental prompt template for subsequent model evaluation.

(Patient information)

- (Basic information)
- - Age: xx years
  - Sex: Male or Female
  - Height: xxx.x kg
  - Weight: xx.x kg
  - Smoking status: Non-smoker, Ex-smoker or Current smoker
  - Drinking pattern: Non-drinker, E-drinker, or Current drinker
  - Alcohol Use Disorders Identification Test (AUDIT) score: (For patients who are current drinkers)

  (Female-specific information)

  - Childbearing potential: Yes or No
  - Pregnancy experience: Yes or No
  - Pregnancy during pregnancy: Yes or No
  - Postpartum depression syndrome: Yes or No
  - Age at menopause: xx years
  - Postmenopausal syndrome: Yes or No
  - Onset of depression at menopause: Yes or No

  (Comorbidities) (All applicable conditions, if any)

  - Allergic/Immunologic disease, Heart disease, Hypertension, Stroke, Respiratory disease, Dermatologic disease, ear , nose and throat (ENT) disease, Endocrine disease, Ophthalmic disease, Gastrointestinal disease, Genitourinary disease, Hematologic cancer, Solid tumor, Musculoskeletal disease, and/or Neurological/Parkinson disease

  (Depression subtype) (All applicable conditions, if any)

  - Anxious, Melancholic, Atypical, or Psychotic

  (Monotherapy and 2-week Response)

  - Main AD (12w): Escitalopram, Paroxetine, Sertraline, Duloxetine, Venlafaxine, Desvenlafaxine, Milnacipran, Miratazapine, Bupropion, Vortioxetine, Tianeptine, or Trazodone
  - Mean dose (12w): xx.x mg - ADT equivalent dose: (12 w): xx.xxx mg
  - Early response at 2 wells (≥20% HAM-D decrease): Yes or No

  (Social-psychological assessments)

  - HAM-D (Hamilton Depression Rating Scale) total score: xx
  - EQ-5D (EuroQol-5 Dimension) index: x.xx
  - SDS (Sheehan Disability Scale) total score: xx
  - PSS (Perceived Stress Scale) total score: xx
  - CD-RISC (Connor-Davidson Resilience Scale) total score: xx
  - MSPSS (Multidimensional Scale of Perceived Social Support) average score: x.xxx

  (Biomarkers)

  - High-sensitivity C-reactive protein (hs-CRP): xxx mg L
  - Tumor necrosis factor-alpha (TNF-α): xx.xx pg/mL
  - Interleukin- 1 beta (IL-1ß): x.xx pg/mL
  - Interleukin-6 (IL-6): x.xxx pg/mL

- Interleukin-4 receptor (I-4R): xxxxx pg/mL
- Interleukin-10 (I-10): xxxxx pg/mL
- Leptin: xx.xx ng/mL
- Ghrelin: xxxxx pg/mL
- Total Cholesterol: xxx mg/dL
- Brain-derived neurotrophic factor (BDNF): xxxx ng/mL

(Mini-International Neuropsychiatric Interview: MINI) (Yes or No)

- Over the past 2 weeks, have you felt depressed or down most of the day, nearly every day?
- Over the past 2 weeks, have you experienced a significantly decreased interest or pleasure in most activities or things you usually enjoy?
- Have you had a nearly daily decrease or increase in appetite, or an unintentional weight loss or gain (±5% of your body weight in 1 month)? If either is Yes, record Yes.
- Have you had insomnia or hypersomnia nearly every day (difficulty falling asleep, trouble staying asleep, early morning awakening, or sleeping too much)?
- Have you spoken or moved more slowly than usual, or have you felt restless or unable to sit still nearly every day? If either is Yes, record Yes.
- Have you felt fatigue or loss of energy nearly every day?
- Have you felt worthless or guilty nearly every day?
- Have you had difficulty concentrating or making decisions nearly every day?
- Have you had recurrent thoughts of self-harm, suicidal ideation, or a wish for death?

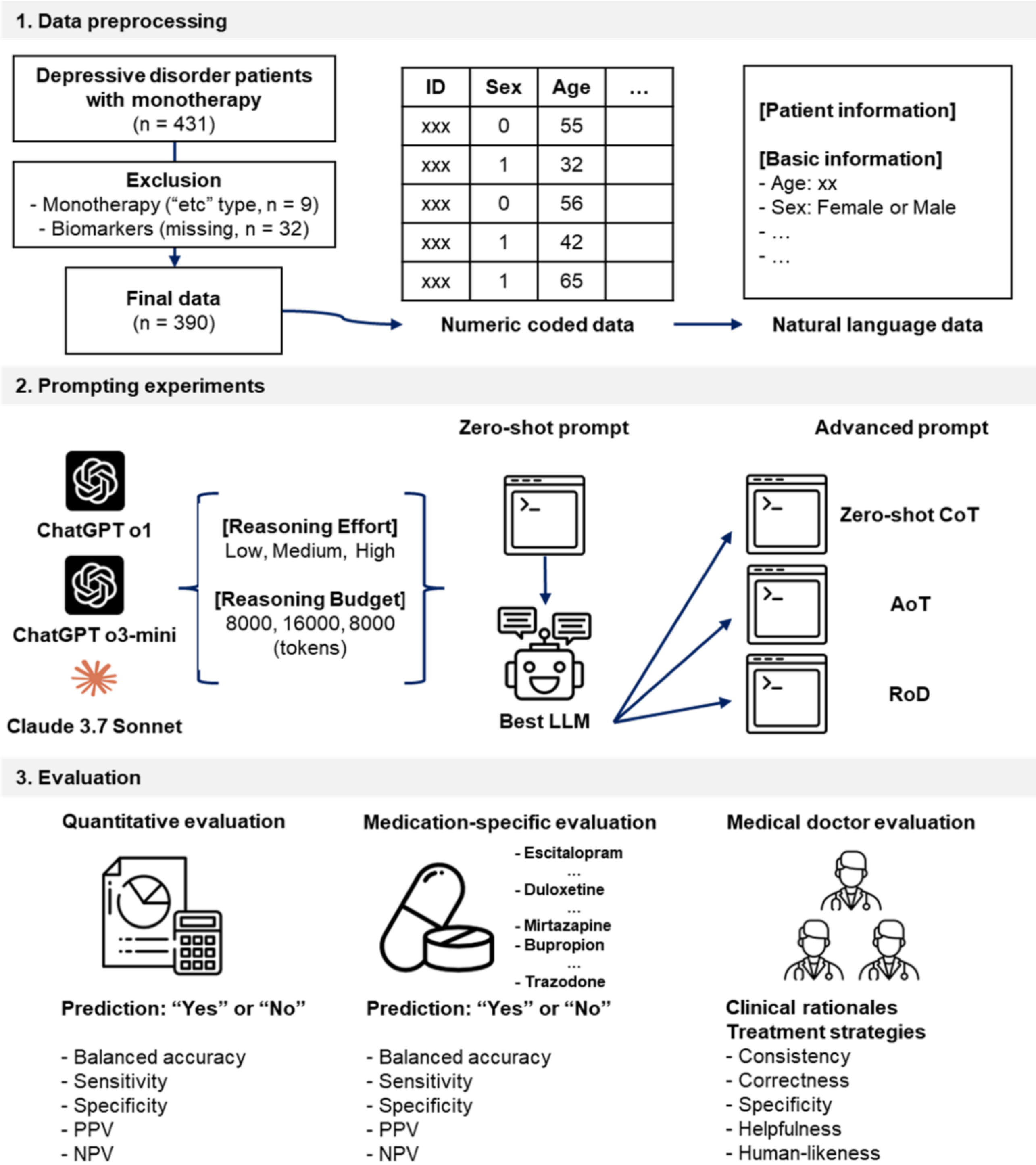
Ethical Considerations

This study was approved by the Chonnam National University Hospital Institutional Review Board (CNUH 2012 - 014). Written informed consent was obtained from all participants. For minors, parental permission and child assent would have been required under institutional and national regulations; however, no minors were enrolled in this study.

Study Design and Zero-Shot Prompting

This study follows the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis guidelines. The design flow is illustrated in Figure 1. Initially, we conducted data preprocessing to prepare input for the LLMs. Subsequently, we used 3 reasoning-based LLMs, including ChatGPT o1 and o3-mini (OpenAI) and Claude 3.7 Sonnet (Anthropic), via an application programming interface to predict 12-week remission in patients with depressive disorder, generating clinical rationales for each prediction and treatment strategies for patients anticipated to not achieve remission; each output consisted of 5 distinct sentences.

Figure 1. Methodological framework for LLM-based prediction of 12-wk remission in patients with depressive disorder. This figure depicts the three-phase methodological approach used in this study: (1) data preprocessing of depressive disorder patients with monotherapy (n=390), including transformation from numeric coded data to natural language format; (2) prompting experiment design; and (3) a comprehensive evaluation framework encompassing quantitative, medication-specific, and clinical assessments. AoT: atom-of-thoughts; CoT: chain-of-thoughts; LLM: large language model; NPV: negative predictive value; PPV: positive predictive value; RoD: referencing of deep research.



We conducted zero-shot experiments to assess the performance of these LLMs. OpenAI's models were evaluated across 3 levels of "reasoning effort" parameters (low, medium, and high), while the Anthropic model was tested at 3 reasoning budget token settings (8000; 16,000; and 32,000 tokens). The detailed structure of the zero-shot prompt is illustrated in Textbox 2.

Textbox 2. Structure of zero-shot prompt. The prompt message remained consistent across all experiments, with only the (patient information) section being systematically replaced with individual patient data for each experimental case.

(Zero-shot prompt)

You are an experienced psychiatrist specializing in depressive disorder. You can access a depressive disorder patient's baseline data, including monotherapy prescribing information and 2-week response.

Your task:

1. Predict the depressive disorder patient's 12-week remission as “Yes” or “No.”
2. Provide a “Clinical Rationale” of exactly five sentences (1~5).
3. If you predict “No,” also provide the next “Treatment Strategy” of exactly five sentences (1~5).
4. Final Output Format (follow precisely):

Remission prediction <Yes or No>

Clinical Rationale:

1. ...
2. ...
3. ...
4. ...
5. ...

Treatment Strategy (only if you predict “No”)

1. ...
2. ...
3. ...
4. ...
5. ...

Below is the patient's baseline data, including (Basic Information), (including (Female-specific Information) if the patient is female), (Comorbidities), (Mini-International Neuropsychiatric Interview), (Depression Subtype) if present, (Adverse Childhood Experiences (ACEs)) if present, (Depression History & Suicidality), (Monotherapy & 2-week Response), (Social-Psychological Assessments), and (Biomarkers).

Please use this data to predict the 12-week remission status (Yes/No) and follow the instructions above.

(Patient Information):

The best-performing zero-shot model, based on balanced accuracy, was further evaluated using advanced prompting strategies to enhance reasoning and interpretability. Specifically, the zero-shot chain-of-thought (CoT) prompting method [31] and the atom-of-thoughts (AoT) technique [32], both of which have shown strong performance on benchmark datasets, were adapted for this study. We also introduced a novel “referencing of deep research (RoD)” prompting strategy, which leverages OpenAI’s deep research [33] to generate research reports that are subsequently incorporated into the zero-shot prompt for additional context.

Finally, our evaluation process comprised multiple sequential phases. First, we conducted a comprehensive quantitative assessment of the zero-shot prompting approaches. Subsequently, using the best-performing model identified through this initial evaluation, we implemented the advanced prompting experiments and subjected them to identical quantitative evaluation methodologies. For the best advanced prompting model, we then performed medication-specific evaluations. Additionally, board-certified medical doctors evaluated the model-generated rationales and treatment strategies.

Advanced Prompting

The zero-shot CoT was implemented by inserting the phrase “Let’s think step by step” immediately before the patient information section in the original zero-shot prompt.

The AoT technique comprised three phases: decomposition (breaking the question into subquestions), contraction (consolidating into an “atomic” question), and final solve (generating predictions with rationale, and, if necessary, subsequent treatment strategies). Detailed prompts appear in Multimedia Appendix 2.

Furthermore, we introduced an RoD prompt to incorporate external evidence regarding 12-week remission outcomes. First, we used OpenAI’s deep research [33] to investigate prior studies on the relationship between our variables of interest and 12-week remission outcomes. A representative dialogue example demonstrating the generation of “deep research report” is presented in Figure 2.

Figure 2. A representative dialogue example of "deep research report" generation. This figure illustrates an example dialogue where the user uses OpenAI's deep research functionality to generate a "deep research report." MAKE BETTER study: MAKE Biomarker Discovery for Enhancing Antidepressant Treatment Effect and Response study.

(User)

You are a highly trained psychiatric research assistant.
Your goal is to investigate peer-reviewed journal articles about 12–week remission outcomes in depressive disorder patients receiving monotherapy.

The following types of monotherapy medications may be considered escitalopram, paroxetine, sertraline, duloxetine, venlafaxine, desvenlafaxine, milnacipran, mirtazapine, bupropion, vortioxetine, tianeptine, or trazodone
You will focus on the following "basic information" variables and their possible relationship to remission rates:

- Age
- Sex
- Height
- Weight
- Employment status
- Living alone
- Smoking status
- Drinking pattern
- Audit (alcohol use disorder test)

You must:
1. Only consider peer-reviewed journal articles.
2. Exclude any articles by authors involved in MAKE BETTER dataset.
3. Summarize your findings for each variable under the specified headings below.
4. If no evidence is found for a particular variable, write "None."
5. Provide a brief "Conclusion" section at the end, summarizing your overall findings.

Use the exact report format shown below, substituting only the bullet points with your findings or "None." Then add the final "Conclusion" after these bullet points.

[Basic Information]
- Age:
- Sex:
- Employment status:
- Living alone:
- Smoking status:
- Drinking pattern:
- Audit: Conclusion:

Conclusion:

(OpenAI's deep research)

Could you specify a preferred time range for the articles (e.g., past 10 years, past 20 years)? Additionally, do you have any preferred databases (e.g., PubMed, PsycINFO) or should I search across all available peer-reviewed sources?

[User]

last 10 years, any databases

(OpenAI's Deep Research)

I will conduct a review on 12–week remission outcomes in depressive disorder patients receiving monotherapy with the specified medications. I will focus on peer–reviewed journal articles from the last 10 years while excluding works by authors involved in MAKE BETTER dataset. The analysis will summarize findings related to the variables: age, sex, employment status, living alone, smoking status, drinking pattern, and alcohol use disorder test (AUDIT). I will let you know when the review is complete.

(OpenAI's deep research)

12–Week Remission in depressive disorder: influence of Demographic and Lifestyle Factors (2015–2025)

Age

Most recent studies do not find a strong or consistent effect of patient age on 12–week remission rates with antidepressant monotherapy. For example, a large trial in a Veterans Affairs population found no significant difference in remission outcomes across age groupspsychiatryonline.org. Earlier analyses (e.g., STAR*D) had suggested that younger patients might attain remission more readily than older patients, but this appears to be confounded by other factors (older individuals often have more chronic illness or comorbid conditions that can hinder response) pmc.ncbi.nlm.nih.gov. In summary, age alone has not shown a clear predictive relationship with 12-week remission in recent literature.

Sex

Findings on sex differences in remission are mixed. Some evidence indicates female patients may experience slightly higher remission rates. A recent large–scale analysis reported that women showed greater antidepressant treatment response than men at all ages pmc.ncbi.nlm.nih.gov. Similarly, the STAR*D trial found being female was associated with higher odds of remission psychiatryonline.org. However, not all studies replicate this advantage–for instance, the VAST–D trial (in a predominantly male sample) found no significant remission difference between sexes psychiatryonline.org. Overall, sex has not been a uniformly reliable predictor of 12–week remission, though some datasets suggest women may respond somewhat better.

Employment status

Employment status is consistently associated with higher remission rates. Multiple studies identify being employed as a favorable prognostic factor. In the STAR*D study, employment was one of the baseline characteristics linked to increased chances of remissionpsychiatryonline.org. More recent analyses reinforce this pattern–for example, an evaluation of depression treatment outcomes noted that employed patients and those with higher socioeconomic status tend to have better antidepressant responses pmc.ncbi.nlm.nih.gov. In a 12–week trial of next-step therapies, remission was significantly more likely in individuals who were working versus unemployed pubmed.ncbi.nlm.nih.gov. Thus, being employed at baseline has emerged as a positive predictor of achieving remission by 12 weeks.

Living alone

None. (No peer–reviewed studies in the last decade were found that directly examine living arrangement–e.g. living alone vs. with others–as a predictor of 12–week remission. Indirectly, low social support has been linked to worse short–term depression outcomes, but no clear evidence addresses living alone per se.)

Smoking status

None. (The impact of smoking on antidepressant remission is not well-studied. A review noted it remains unclear whether smokers differ from non–smokers in depression treatment outcomes pmc.ncbi.nlm.nih.gov. No consistent evidence links baseline smoking status to 12–week remission rates in depressive disorder monotherapy trials.)

Drinking pattern

None. (Limited data exist on how baseline drinking habits affect remission. Patients with severe alcohol use are often excluded from trials. A recent analysis found that comorbid alcohol use disorder did not significantly alter 12–week remission rates pubmed.ncbi.nlm.nih.gov. Overall, no clear evidence ties drinking frequency or quantity to remission likelihood.)

AUDIT score

None. (No studies were identified that specifically evaluated the Alcohol Use Disorder Identification Test (AUDIT) scores as predictors of 12–week remission. This specific alcohol-use measure has not been reported as an influential factor in recent antidepressant trials.)

Conclusion

In summary, among the patient variables reviewed, employment status shows the most consistent relationship with 12–week remission–employed patients tend to have better outcomes. Sex differences are modest: some studies suggest women have a slight advantage in achieving remission, but this finding is not universal. Age does not appear to be a decisive factor for remission in recent evidence once other factors are considered. For other variables–living situation (living alone), smoking, and alcohol use (including AUDIT scores)–the current literature reveals no clear associations with 12–week remission rates (indeed, little to no direct evidence was found for these). Overall, apart from socioeconomic factors like employment, demographic and lifestyle factors have shown only weak or inconsistent links to short-term remission in depressive disorder. Further research may clarify whether subtle effects exist or confirm that these factors are not major determinants of early treatment success.

This systematic examination deliberately excluded publications authored by contributors to the MAKE BETTER dataset to mitigate potential confirmation bias and ensure methodological independence in our analysis framework. We then consolidated these findings into a "deep research report" and subsequently integrated this report into the zero-shot prompt to design the RoD prompt. The prompt for conducting the "deep research report" and the RoD prompt is presented in Textbox 3.

Textbox 3. The prompts of deep research and referencing of deep research. The deep research prompt template, used for OpenAI's deep research functionality, instructs the model to function as a research assistant. The hierarchical structure of the patient information is systematically mapped to the prompt parameters, with bracketed "[]" items from the (patient information) corresponding to (big category) classifications, while hyphenated "-" elements are allocated to (small category) designations. The referencing of the deep research prompt template incorporates outputs from deep research operations into the (deep research report) section, emulating the clinical reasoning process whereby practitioners consult and integrate contemporary research literature before formulating diagnostic conclusions. Sections marked as (omit) indicate portions where identical prompt text from previously described templates has been elided for clarity.

(Deep Research Prompt)

You are a highly trained psychiatric research assistant.

Your goal is to investigate peer-reviewed journal articles about 12-week remission outcomes in depressive disorder patients receiving monotherapy. The following types of monotherapy medications may be considered: escitalopram, paroxetine, sertraline, duloxetine, venlafaxine, desvenlafaxine, milnacipran, mirtazapine, bupropion, vortioxetine, tianeptine, or trazodone.

You will focus on the following "(Big category)" variables and their possible relationship to remission rates:

(Small Category)

- Age, Sex ... (omitted) ... Homocysteine

1. Only consider peer-reviewed journal articles.
2. Exclude any articles by authors involved in the MAKE BETTER dataset.
3. Summarize your findings for each variable under the specified headings below.
4. If no evidence is found for a particular variable, write "None."
5. Provide a brief "Conclusion" section at the end, summarizing your overall findings.

Use the exact report format shown below, substituting only the bullet points with your findings or "None." Then add the final "Conclusion" after these bullet points.

(Big Category)

- (Small Category)
- ...

Conclusion:

(RoD prompt)

You are an experienced ... (omitted) ... 2-week response, as well as a deep research report summarizing findings on 12-week remission outcomes for depressive disorder monotherapy.

Reason as needed, incorporating your own expertise and the research evidence contained in the deep research report below.

(Deep research report)

Your task:

1. ... (omitted) ...
2. ... (omitted) ...
3. ... (omitted) ...
4. Do not copy research text verbatim. Summarize relevant parts like a clinician referencing journal articles.
5. Final output format (follow precisely): ... (omitted) ...

The model was instructed to reference rather than directly replicate relevant insights from the "deep research report" when generating predictions and clinical rationales, thereby emulating the manner in which a practicing clinician would consult and synthesize findings from journal articles.

Evaluation

For the 12-week remission prediction task, we designated "yes" as the positive class and "no" as the negative class. We computed balanced accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) to compare quantitative performance. Additionally, to evaluate efficiency, we recorded both the inference generation cost and the average generation time (in seconds). The best-performing zero-shot model was selected based on balanced accuracy, reflecting the equal importance of both classes.

Additionally, we performed benchmarking analyses using logistic regression, random forest, and XGBoost models, evaluated through a patient-level stratified 15% hold-out design with repeated 10×5-fold cross-validation, reporting balanced accuracy, sensitivity, specificity, PPV, and NPV with 95% CIs across random seeds.

Subsequently, we applied the CoT, AoT, and RoD prompting methods to this best-performing model, compared their final performance using the same metrics, and further examined the medication-specific performance of the model that achieved the highest overall balanced accuracy.

Finally, 3 evaluators (2 psychiatry residents with >2 years of training and 1 psychiatrist specializing in depressive disorder with >10 y of experience) independently reviewed the clinical rationales and following treatment strategies generated for the correctly predicted cases by the best-performing model. They assessed these outputs across 5 domains (consistency, correctness, specificity, helpfulness, and human likeness) using a 5-point rating scale [34]. Consistency measured how closely the generated text aligned with the predicted answers, correctness evaluated its medical accuracy, specificity assessed its level of detail, helpfulness examined its clinical use, and human likeness considered how similar it was to typical human judgment.

Results

Baseline Demographics and Clinical Characteristics

Table 1 summarizes the baseline demographics and clinical characteristics across different monotherapy groups. The study population consisted of 244 patients prescribed SSRIs (escitalopram: n=159, 65%; paroxetine: n=60, 25%; and sertraline: n=25, 10%), 33 patients receiving SNRIs (duloxetine: n=20, 61%; venlafaxine: n=10, 30%; desvenlafaxine: n=2, 6%; and milnacipran: n=1, 3%), 99 patients on mirtazapine, 9 patients prescribed bupropion, and 5 patients taking other antidepressants (vortioxetine: n=3, 60%; tianeptine: n=1, 20%; and trazodone: n=1, 20%).

Table . Baseline demographics and clinical characteristics of preprocessed patients with depressive disorder, stratified according to the types of prescribed monotherapy (n=390).

Characteristics	Types of prescribed monotherapy				
	SSRI[a] (n=244)	SNRI[b] (n=33)	Mirtazapine (n=99)	Bupropion (n=9)	Others (n=5)
Sex, n (%)					
Female	175 (72)	25 (76)	77 (78)	4 (44)	4 (80)
Male	69 (28)	8 (24)	22 (22)	5 (56)	1 (20)
Employment status, n (%)					
Yes	180 (74)	25 (76)	67 (68)	7 (78)	4 (80)
No	64 (26)	8 (24)	32 (32)	2 (22)	1 (20)
Living alone, n (%)					
Yes	41 (17)	2 (6)	17 (17)	2 (22)	3 (60)
No	203 (83)	31 (94)	82 (83)	7 (78)	2 (40)
12-week remission, n (%)					
Yes	83 (34)	13 (39)	42 (42)	3 (33)	1 (20)
No	161 (66)	20 (61)	57 (58)	6 (67)	4 (80)
Age (y), mean (SD)	56.8 (14.5)	58.4 (9.5)	60.4 (14.1)	46.4 (14.6)	58.6 (8.0)
Height (cm), mean (SD)	159.9 (8.9)	157.3 (8.1)	159.1 (7.7)	165.1 (6.6)	157.9 (8.7)
Weight (kg), mean (SD)	59.7 (10.5)	58.3 (9.2)	59.4 (9.7)	59.0 (11.8)	60.7 (8.5)
HAM-D[c], mean (SD)	20.4 (4.1)	20.5 (4.1)	21.2 (3.9)	18.6 (4.7)	22.2 (4.4)

[a]SSRI: selective serotonin reuptake inhibitor.

[b]SNRI: serotonin and norepinephrine reuptake inhibitor.

[c]HAM-D: the Hamilton Depression Rating Scale.

Among the total cohort (n=390), female participants constituted the majority (285/390, 73%), with similar gender distribution across the SSRIs (175/244, 72%), SNRIs (25/33, 76%), and mirtazapine groups (77/99, 78%). Employment was reported by 74% (180/244) of SSRI users, 76% (25/33) of SNRI users, and 68% (67/99) of mirtazapine users. At the 12-week assessment, 34% (83/244) of SSRI users, 39% (13/33) of SNRI users, and 42% (42/99) of mirtazapine users achieved remission.

The mean baseline HAM-D scores ranged from 18.6 (SD 4.7) to 22.2 (4.4) points, with participants in the mirtazapine group being slightly older (mean 60.4, SD 14.1 y) than those in the bupropion group (mean 46.4, SD 14.6 y).

Performance of Zero-Shot Prompting

The zero-shot performance section of Table 2 delineates the comparative outcomes of zero-shot experiments conducted with

OpenAI's ChatGPT o1 and o3-mini models across 3 distinct levels of reasoning effort, namely "low," "medium," and "high," as well as for Anthropic's Claude 3.7 Sonnet under 3 varying token budget settings (8000; 16,000; and 32,000 tokens). The findings indicate that all models demonstrated sensitivity values ranging from 0.6690 to 0.9085, suggesting that a significant proportion of patients who achieved remission were accurately identified. Conversely, specificity, which measures the correct identification of patients who did not achieve remission, exhibited lower values, ranging from 0.3185 to 0.6331 across the evaluated LLMs.

Table . Quantitative performance of zero-shot and advanced prompting techniques across 390 samples, including balanced accuracy, sensitivity, specificity, PPV,[a] and NPV[b].

Prompting, models, and reasoning parameters	Balanced accuracy	Sensitivity	Specificity	PPV	NPV	Time per generation (s)	Total cost (US $)
Zero-shot							
ChatGPT o1							
Low	0.6135	0.9085	0.3185	0.4329	0.8587	11.44	22.36
Medium	0.6382	0.9014	0.3750	0.4523	0.8692	19.63	35.20
High	0.6333	0.8592	0.4073	0.4535	0.8347	30.08	53.07
ChatGPT o3-mini							
Low	0.6121	0.8169	0.4073	0.4411	0.7953	4.84	1.14
Medium	0.6091	0.8028	0.4153	0.4402	0.7863	8.89	2.00
High	0.6323	0.8169	0.4476	0.4585	0.8102	20.43	4.39
Claude 3.7 Sonnet							
8000	0.6349	0.6972	0.5726	0.4829	0.7676	22.23	9.81
16,000	0.6511	0.6690	0.6331	0.5108	0.7696	23.78	10.90
32,000	0.6656	0.7183	0.6129	0.5152	0.7917	26.84	11.58
Zero-shot CoT[c]							
Claude 3.7 Sonnet with 32,000 tokens	0.6319	0.6549	0.6089	0.4895	0.7550	27.24	12.13
Zero-shot AoT[d]							
Claude 3.7 Sonnet with 32,000 tokens	0.6522	0.4859	0.8185	0.6053	0.7355	126.92	57.56
Zero-shot RoD[e]							
Claude 3.7 Sonnet with 32,000 tokens	0.6697	0.7183	0.6210	0.5204	0.7938	43.88	39.56

[a]PPV: positive predictive value.
[b]NPV: negative predictive value.
[c]CoT: chain-of-thoughts.
[d]AoT: atom-of-thoughts.
[e]RoD: referencing of deep research.

As the reasoning effort increased, all 3 models showed enhancements in both specificity and balanced accuracy. Specifically, the ChatGPT o1 model's specificity improved from 0.3185 to 0.4073, with balanced accuracy rising from 0.6135 to 0.6333. Similarly, the ChatGPT o3-mini model experienced an increase in specificity from 0.4073 to 0.4476, alongside an improvement in balanced accuracy from 0.6121 to 0.6323. The Claude 3.7 Sonnet model also demonstrated an increase in specificity from 0.5726 to 0.6129, with a modest rise in balanced accuracy from 0.6349 to 0.6656.

From a computational efficiency standpoint, an increase in reasoning level generally resulted in heightened time and cost requirements across all models. Across all models evaluated, ChatGPT o1 incurred the highest overall costs, with total expenses ranging from $22.36 to $53.07. In contrast, ChatGPT o3-mini emerged as the most cost-effective option, with total costs between $1.14 and $4.39, rendering it the least expensive model. Furthermore, ChatGPT o3-mini exhibited superior speed efficiency, with task completion times ranging from 4.84 to 20.43 seconds, outperforming the other models in computational efficiency.

Conversely, Claude 3.7 Sonnet maintained a relatively stable computational profile across varying token budgets, with task completion times ranging from 22.23 seconds at the 8000-token setting to 26.84 seconds at the 32,000-token setting, and total costs increasing modestly from $9.81 to $11.58. Despite requiring more time per task than ChatGPT o3-mini at lower settings, Claude 3.7 Sonnet's costs remained significantly lower than those of ChatGPT o1 at higher reasoning levels, while achieving the best overall performance, as evidenced by its balanced accuracy of 0.6656 at the 32,000-token reasoning budget. The detailed confusion matrices for all zero-shot prompting experiments are presented in Multimedia Appendix 3.

Performance of Advanced Prompting

The advanced prompting (zero-shot CoT, AoT, and RoD) performance section of Table 2 outlines the performance metrics of 3 advanced prompt strategies applied to the Claude 3.7 Sonnet model using a 32,000-token reasoning budget, which demonstrated the best performance in the zero-shot context.

Among the advanced prompt strategies, the zero-shot CoT exhibited a balanced accuracy of 0.6319, with sensitivity and specificity values of 0.6549 and 0.6089, respectively, alongside a PPV of 0.4895 and an NPV of 0.7550. This performance is marginally lower than that of Claude 3.7 Sonnet's zero-shot approach, particularly in terms of sensitivity and balanced accuracy.

The AoT strategy demonstrated a balanced accuracy of 0.6522, with a sensitivity of 0.4859 and a specificity of 0.8185. Its PPV and NPV were recorded at 0.6053 and 0.7355, respectively, while the time per task reached 126.92 seconds, and total costs escalated to $57.56, indicating a significant increase in computational resource demands compared to the zero-shot approach of Claude 3.7 Sonnet.

In contrast, the RoD approach achieved the highest balanced accuracy among the advanced prompts at 0.6697, with a sensitivity of 0.7183 and a specificity of 0.6210, slightly surpassing the performance of Claude 3.7 Sonnet's zero-shot method. However, RoD's time per task was approximately 1.63 times greater, and its total cost was approximately 3.42 times that of the zero-shot setting. The detailed confusion matrices for all advanced prompting experiments are presented in Multimedia Appendix 4.

For reference, conventional ML models trained on the numerically coded dataset achieved balanced accuracies ranging from 0.6077 to 0.7371 and sensitivities from 0.3533 to 0.6364 with overlapping 95% CIs (Multimedia Appendix 5).

Medication-Specific Performance

Table 3 presents the performance metrics for the RoD strategy across various antidepressants, including SSRIs (escitalopram, paroxetine, and sertraline), SNRIs (duloxetine, venlafaxine, desvenlafaxine, and milnacipran), mirtazapine, bupropion, and others (vortioxetine, tianeptine, and trazodone), along with the number of correct predictions for both remission and nonremission outcomes. Among antidepressants with more than 50 cases, escitalopram (n=159), mirtazapine (n=99), and paroxetine (n=60) achieved balanced accuracies of 0.6799, 0.6873, and 0.6375, respectively.

Table . Quantitative performance of RoD[a] prompting by medications, applied to Claude 3.7 Sonnet configured with 32,000 reasoning budget tokens.

Medications	Balanced accuracy	Sensitivity	Specificity	PPV[b]	NPV[c]	Correct predictions (yes), n/N	Correct predictions (no), n/N
SSRI[d]							
Escitalopram	0.6799	0.7407	0.6190	0.5000	0.8228	40/54	65/105
Paroxetine	0.6375	0.8000	0.4750	0.4324	0.8261	16/20	19/40
Sertraline	0.7083	0.6667	0.7500	0.6000	0.8000	6/9	12/16
SNRI[e]							
Duloxetine	0.6190	0.6667	0.5714	0.4000	0.8000	4/6	8/14
Venlafaxine	0.7083	0.7500	0.6667	0.6000	0.8000	3/4	4/6
Desvenlafaxine	0.5000	1.0000	0.0000	1.0000	0.0000	2/2	0/0
Milnacipran	0.0000	0.0000	0.0000	0.0000	0.0000	0/1	0/0
Mirtazapine	0.6873	0.6905	0.6842	0.6170	0.7500	29/42	39/57
Bupropion	0.7500	0.6667	0.8333	0.6667	0.8333	2/3	5/6
Others							
Vortioxetine	0.0000	0.0000	0.0000	0.0000	0.0000	0/1	0/2
Tianeptine	0.5000	0.0000	1.0000	0.0000	1.0000	0/0	1/1
Trazodone	0.5000	0.0000	1.0000	0.0000	1.0000	0/0	1/1

[a]RoD: referencing of deep research.
[b]PPV: positive predictive value.
[c]NPV: negative predictive value.
[d]SSRI: selective serotonin reuptake inhibitor.
[e]SNRI: serotonin and norepinephrine reuptake inhibitor.

Medical Doctor Evaluation of Model-Generated Rationales and Treatment Strategies

A total of 3 clinical evaluators independently assessed the clinical rationales and treatment strategies generated by the best-performing model for 256 correctly predicted cases. As presented in Table 4, the highest total rating was observed for correctness (mean, 4.3, SD 0.7). Consistency, specificity, and helpfulness also received favorable evaluations (means 4.2, 4.2, and 4.2, respectively). Human likeness received the lowest but still positive rating (mean 3.6, SD 1.7). Notably, the board-certified psychiatrist rated helpfulness highest (mean 4.5, SD 0.6), while consistency scores varied most between evaluators, ranging from a mean of 3.4 to 4.9. To demonstrate the interpretability of the model's reasoning process, one representative remission case ("yes") and one nonremission case ("no") were selected as examples, each accompanied by psychiatrist evaluations and comments. These illustrative cases are presented in Multimedia Appendices 6 and 7.

Table . Evaluations on clinical rationales and treatment strategies assigned by a board-certified psychiatrist and psychiatry residents for the clinical outputs produced by the best model across 256 correctly predicted cases.[a]

	Consistency, mean (SD)	Correctness, mean (SD)	Specificity, mean (SD)	Helpfulness, mean (SD)	Human likeness, mean (SD)
Psychiatrist	3.4 (0.6)	4.3 (0.5)	4.0 (0.5)	4.5 (0.6)	3.5 (0.5)
Resident 1	4.3 (0.5)	4.4 (0.7)	4.2 (0.6)	4.3 (0.7)	3.9 (2.6)
Resident 2	4.9 (0.4)	4.2 (0.8)	4.3 (0.8)	3.9 (1.3)	3.4 (1.2)
Residents[b]	4.6 (0.5)	4.3 (0.7)	4.3 (0.7)	4.1 (1.1)	3.6 (2.0)
Total[c]	4.2 (0.8)	4.3 (0.7)	4.2 (0.7)	4.2 (1.0)	3.6 (1.7)

[a]Assessments were conducted across 5 domains using a 5-point scale (1-5), with higher scores indicating better performance.
[b]The "residents" row represents the aggregated scores from both residents.
[c]"Total" indicates the combined assessment across all 3 evaluators.

Discussion

Principal Findings

Reasoning-based LLMs, especially when guided by research-informed prompting strategies, demonstrate promising potential in predicting antidepressant treatment response among patients with depressive disorder. To the best of our knowledge, this is among the first applications of LLMs for forecasting remission outcomes in depression, extending beyond prior approaches that primarily used traditional statistical and ML models [7-9,35,36].

In zero-shot contexts, all models showed higher sensitivity (0.6690 - 0.9085) than specificity (0.3185 - 0.6331). Balanced accuracy improved with enhanced reasoning: ChatGPT o1 by 3.22%, ChatGPT o3-mini by 3.3%, and Claude 3.7 Sonnet by 4.8%, with Claude achieving the highest performance (0.6656) at 32,000 budget tokens. This supports prior findings on reasoning capabilities' importance in medical applications [37,38], suggesting that enhanced reasoning depth improves LLM performance in specific clinical tasks. Moreover, our proposed RoD technique, which emulates how clinicians incorporate contemporary research findings into their clinical reasoning process, outperformed zero-shot CoT and AoT with highest balanced accuracy (0.6697). While requiring further research, RoD appears effective for psychiatric prediction tasks. Compared with conventional ML baselines (Multimedia Appendix 5), which achieved balanced accuracies of 0.6077 to 0.7371 and sensitivities of 0.3533 to 0.6364, our reasoning-based LLM approach demonstrated higher sensitivity, indicating improved identification of patients who ultimately achieved remission. Analyzing medication-specific performance after excluding antidepressants with fewer than 10 cases, NPV remained high (>0.75) across all medications. For escitalopram, which was the most frequently prescribed antidepressant in the cohort (n=159), the RoD prompting approach achieved a balanced accuracy of 0.6799. Although direct comparison is limited by differences in sample size and methodology, this value is numerically higher than the 0.61 balanced accuracy reported in a prior partial least squares regression analysis of 92 escitalopram-treated patients [36], suggesting that reasoning-based LLMs may achieve comparable or potentially improved predictive capability within a single antidepressant group.

A particularly noteworthy finding is the contrasting performance between traditional reasoning approaches (CoT/AoT) and our knowledge-augmented RoD strategy. While CoT and AoT showed minimal improvement or even slight performance degradation compared to zero-shot prompting, RoD achieved consistent improvements across all metrics. This divergence suggests that for clinical pattern-recognition tasks, the decomposition of reasoning steps alone (as in CoT/AoT) may introduce unnecessary complexity without meaningful benefit. In contrast, RoD's incorporation of synthesized research evidence appears to provide crucial contextual priors that enhance prediction accuracy. This mirrors actual clinical practice, where psychiatrists integrate empirical evidence from literature with patient-specific data rather than relying solely on sequential logical reasoning.

The superior performance of RoD likely stems from its ability to leverage documented patterns in depressive disorder treatment outcomes, effectively providing the model with a knowledge base of established clinical associations. This approach compensates for the inherent limitations of LLMs in medical domains, where training data may not adequately capture the full spectrum of clinical scenarios. Furthermore, by grounding predictions in research evidence, RoD may reduce the risk of hallucinations or clinically implausible outputs that can occur with pure reasoning approaches, a critical concern in medical artificial intelligence (AI) applications [37]. These findings align with recent evidence suggesting that retrieval-augmented approaches enhance LLM reliability in clinical contexts [38]. The hybrid strategy combining LLM reasoning with structured knowledge integration may represent an optimal approach for clinical prediction tasks, particularly in psychiatry, where outcomes are influenced by complex biopsychosocial factors [39].

Clinical Implications

Clinical evaluation of the model-generated rationales and treatment suggestions revealed high ratings for correctness, consistency, specificity, and perceived helpfulness, indicating that reasoning-based LLMs can produce clinically coherent and contextually relevant outputs. Favorable assessments by practicing clinicians further suggest their potential as valuable adjuncts in real-world clinical decision-making, particularly for the early identification of patients at risk of treatment nonremission. Unlike prior models focused mainly on predictive performance, our approach emphasizes interpretability and clinician usability, which are key elements for real-world application. By integrating biomarker and clinical data with advanced reasoning, LLMs may support more personalized and effective treatment decisions. Nonetheless, relatively lower ratings for human likeness highlight the need for improved communication style to foster trust and interpretability in clinical practice.

The high NPV (>0.75) across all medication classes suggests particular utility as a screening tool to identify patients unlikely to achieve remission with standard first-line treatments. This could enable a stratified care approach, where predicted nonresponders receive enhanced monitoring, earlier treatment adjustments, or augmentation strategies, potentially reducing the typical 12-week trial-and-error period. Such implementation aligns with recent frameworks for integrating AI into clinical psychiatry that emphasize augmentation rather than replacement of clinical judgment [40]. The RoD prompting strategy required an average processing time of 43.88 seconds per patient, suggesting that real-time clinical application is feasible within standard consultation time frames.

From a health economics perspective, early identification of nonresponders could substantially reduce costs associated with prolonged ineffective treatments, emergency interventions, and productivity losses. The ability to provide detailed clinical rationales distinguishes our approach from black-box algorithms, addressing a critical barrier to AI adoption in psychiatry, where understanding the reasoning behind recommendations is essential for clinical acceptance and regulatory approval [41]. Moreover, the cloud-based nature of LLMs enables deployment without specialized hardware, making this technology accessible to resource-limited settings where psychiatric expertise may be scarce [42].

Successful clinical implementation would require integration with electronic health records, development of user-friendly interfaces, and establishment of clear protocols for acting on model predictions. The model's ability to suggest alternative treatment strategies when predicting nonremission provides actionable guidance rather than mere risk stratification, potentially improving clinical utility. Furthermore, the transparent reasoning process could serve an educational function, helping less experienced clinicians understand factors influencing treatment response and potentially improving their clinical reasoning skills over time [43]. Prospective validation studies are warranted to confirm these findings in real-world clinical settings.

Limitations

Despite promising findings, several limitations warrant consideration. First, while our approach demonstrated robust sensitivity (0.7183) and NPV (0.7938), the relatively low PPV (0.5204) may generate false positives, potentially complicating treatment planning for patients misclassified as achieving remission [44]. The relatively modest PPV observed in our model should be interpreted in light of the low remission prevalence in our cohort, a condition known to constrain PPV despite adequate discriminative performance. Although PPV was modest, the model demonstrated balanced accuracy and sensitivity at clinically meaningful levels, supporting its capacity for reliable risk stratification in a heterogeneous depressive population. Importantly, the high NPV suggests that the model may be particularly effective for identifying patients unlikely to achieve remission, thereby enabling early treatment modifications or augmentation strategies to improve outcomes. These findings emphasize that the model is intended as an adjunctive decision-support tool, and its predictions should be integrated with comprehensive clinical assessments.

Medication-specific analyses revealed sample imbalances (Table 3), with escitalopram dominating (n=159) and several medications having fewer than 20 cases. Although overall model performance remained robust, medication-specific metrics should be interpreted with caution for drugs with limited samples. This imbalance reflects real-world prescribing patterns but limits our ability to make definitive conclusions about model performance for less commonly prescribed antidepressants [45]. Future studies should either focus on medications with adequate sample sizes or use targeted recruitment strategies to ensure sufficient representation across all medication classes.

Our clinical evaluation methodology has notable limitations. The assessment was conducted by only 3 evaluators from a single institution, potentially introducing institutional bias and limiting generalizability. More critically, evaluation was restricted to correctly predicted cases, which likely inflates perceived quality scores and fails to capture model behavior in misclassification scenarios. Future studies should incorporate multi-institutional evaluators and a comprehensive assessment of both correct and incorrect predictions to provide more robust validation of AI-assisted diagnostic approaches.

Finally, the RoD method requires further comparative evaluation against alternative knowledge-augmented techniques to determine its optimal application in psychiatric contexts. Validation in ethnically diverse populations with larger numbers of clinical expert appraisals remains essential. Prospective randomized trials are needed to evaluate whether model recommendations improve clinical outcomes and decision-making in practice.

Conclusions

In conclusion, this study demonstrates the promising potential of reasoning-based LLMs for predicting antidepressant treatment response in patients with depressive disorder. Our findings highlight the superior performance of the RoD technique, which achieved the highest performance by integrating research evidence with clinical reasoning, representing an important advance toward AI-assisted clinical decision support in psychiatry. The high NPV (>0.75) across medications suggests particular use as a screening tool for identifying patients unlikely to achieve remission with standard treatments. While limitations exist, including the need for validation in diverse populations and larger-scale clinical evaluations, the positive assessment by clinical experts validates the potential use of these approaches. Future research should focus on expanding real-world treatment outcome datasets, conducting multi-institutional clinical evaluations, and developing models that can predict both the magnitude of treatment response and suggest personalized next-step strategies. These advances could enable clinicians to make more informed, evidence-based decisions in selecting the most effective personalized treatment strategies for patients with depressive disorder.

Funding

This research was supported by the Bio & Medical Technology Development Program and of the National Research Foundation funded by the Korean government (Ministry of Science and ICT) (RS-2024-00440371 and RS-2024-00457381).

Data Availability

The data that support the findings of the study are available from Jae-Min Kim upon reasonable request.

Authors' Contributions

JMK had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis
Concept and design: JMK, HL, HJK, JHP
Acquisition, analysis, and interpretation of data: All authors
Manuscript drafting: JMK, HL, HJK, JHP
Critical review of the manuscript for important intellectual content: All authors
Statistical analysis: JMK, HL, HJK, JHP
Funding: JMK and HL
Administrative, technical, or material support: JMK, HJK, JHJ, SGK, JWK
Supervision: All authors

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials on the MAKE BETTER study.
[DOCX File, 26 KB - mental_v13i1e83352_app1.docx]

Multimedia Appendix 2
Structure of the atom-of-thoughts (AoT) prompt.
[DOCX File, 409 KB - mental_v13i1e83352_app2.docx]

Multimedia Appendix 3
Confusion matrices for each zero-shot prompting under varying reasoning levels or token budgets.
[DOCX File, 294 KB - mental_v13i1e83352_app3.docx]

Multimedia Appendix 4
Confusion matrices for advanced prompting strategies.
[DOCX File, 108 KB - mental_v13i1e83352_app4.docx]

Multimedia Appendix 5
Predictive performance of machine learning models for 12-week remission classification.
[DOCX File, 19 KB - mental_v13i1e83352_app5.docx]

Multimedia Appendix 6
Representative remission ("yes") case generated by the RoD prompting strategy.
[DOCX File, 1046 KB - mental_v13i1e83352_app6.docx]

Multimedia Appendix 7
Representative remission ("no") case generated by the RoD prompting strategy.
[DOCX File, 1417 KB - mental_v13i1e83352_app7.docx]

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150. [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
2. Kim JM, Kim SW, Stewart R, et al. Predictors of 12-week remission in a nationwide cohort of people with depressive disorders: the CRESCEND study. Hum Psychopharmacol 2011 Jan;26(1):41-50. [doi: 10.1002/hup.1168] [Medline: 21344501]
3. Jin YT, Kim HY, Jhon M, et al. Prediction of 12-Week remission by psychopharmacological treatment step in patients with depressive disorders. Psychiatry Investig 2022 Oct;19(10):866-871. [doi: 10.30773/pi.2022.0160] [Medline: 36327967]
4. Walter HJ, Abright AR, Bukstein OG, et al. Clinical practice guideline for the assessment and treatment of children and adolescents with major and persistent depressive disorders. J Am Acad Child Adolesc Psychiatry 2023 May;62(5):479-502. [doi: 10.1016/j.jaac.2022.10.001] [Medline: 36273673]

5. Perlman K, Benrimoh D, Israel S, et al. A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. J Affect Disord 2019 Jan 15;243:503-515. [doi: 10.1016/j.jad.2018.09.067] [Medline: 30286415]
6. Sharma A, Barrett MS, Cucchiara AJ, Gooneratne NS, Thase ME. A breathing-based meditation intervention for patients with major depressive disorder following inadequate response to antidepressants: a randomized pilot study. J Clin Psychiatry 2017 Jan;78(1):e59-e63. [doi: 10.4088/JCP.16m10819] [Medline: 27898207]
7. Benoit JRA, Dursun SM, Greiner R, et al. Using machine learning to predict remission in patients with major depressive disorder treated with desvenlafaxine. Can J Psychiatry 2022 Jan;67(1):39-47. [doi: 10.1177/07067437211037141] [Medline: 34379019]
8. Salem H, Huynh T, Topolski N, et al. Temporal multi-step predictive modeling of remission in major depressive disorder using early stage treatment data; STAR*D based machine learning approach. J Affect Disord 2023 Mar 1;324:286-293. [doi: 10.1016/j.jad.2022.12.076] [Medline: 36584711]
9. Carr E, Rietschel M, Mors O, et al. Optimizing the prediction of depression remission: a longitudinal machine learning approach. Am J Med Genet B Neuropsychiatr Genet 2025 Apr;198(3):e33014. [doi: 10.1002/ajmg.b.33014] [Medline: 39470297]
10. Zhukovsky P, Trivedi MH, Weissman M, Parsey R, Kennedy S, Pizzagalli DA. Generalizability of treatment outcome prediction across antidepressant treatment trials in depression. JAMA Netw Open 2025 Mar 3;8(3):e251310. [doi: 10.1001/jamanetworkopen.2025.1310] [Medline: 40111362]
11. Cheng SW, Chang CW, Chang WJ, et al. The now and future of ChatGPT and GPT in psychiatry. Psychiatry Clin Neurosci 2023 Nov;77(11):592-596. [doi: 10.1111/pcn.13588] [Medline: 37612880]
12. Shin D, Kim H, Lee S, Cho Y, Jung W. Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: instrument validation study. J Med Internet Res 2024 Sep 18;26:e54617. [doi: 10.2196/54617] [Medline: 39292502]
13. Omar M, Soffer S, Charney AW, Landi I, Nadkarni GN, Klang E. Applications of large language models in psychiatry: a systematic review. Front Psychiatry 2024;15:1422807. [doi: 10.3389/fpsyt.2024.1422807] [Medline: 38979501]
14. Introducing openai o1-preview. OpenAI. 2024. URL: https://openai.com/research/introducing-openai-o1-preview [accessed 2025-12-23]
15. OpenAI o3-mini: pushing the frontier of cost-effective reasoning. OpenAI. 2025. URL: https://openai.com/research/openai-o3-mini [accessed 2025-12-23]
16. Claude's extended thinking. Anthropic. 2025. URL: https://www.anthropic.com/news/visible-extended-thinking [accessed 2025-12-23]
17. Xie Y, Wu J, Tu H, Yang S, Zhao B, Zong Y, et al. A preliminary study of o1 in medicine: are we closer to an ai doctor? arXiv. Preprint posted online on Sep 23, 2024. [doi: 10.48550/arXiv.2409.15277]
18. Chen J, Cai Z, Ji K, Wang X, Liu W, Wang R, et al. Huatuogpt-o1, towards medical complex reasoning with llms. arXiv. Preprint posted online on Dec 25, 2024. [doi: 10.48550/arXiv.2412.18925]
19. Mondillo G, Colosimo S, Perrotta A, Frattolillo V, Masino M. Comparative evaluation of advanced AI reasoning models in pediatric clinical decision support: chatgpt O1 vs. deepseek-r1. medRxiv. Preprint posted online on Jan 27, 2025. [doi: 10.1101/2025.01.27.25321169]
20. Mondillo G, Masino M, Colosimo S, Perrotta A, Frattolillo V. Evaluating AI reasoning models in pediatric medicine: a comparative analysis of o3-mini and o3-mini-high. medRxiv. Preprint posted online on Feb 27, 2025. [doi: 10.1101/2025.02.27.25323028]
21. Xu P, Wu Y, Jin K, Chen X, He M, Shi D. DeepSeek-R1 outperforms Gemini 2.0 Pro, OpenAI o1, and o3-mini in bilingual complex ophthalmology reasoning. Adv Ophthalmol Pract Res 2025;5(3):189-195. [doi: 10.1016/j.aopr.2025.05.001] [Medline: 40678192]
22. Kang HJ, Kim JW, Kim SY, et al. The MAKE biomarker discovery for enhancing antidepressant treatment effect and response (MAKE BETTER) study: design and methodology. Psychiatry Investig 2018 May;15(5):538-545. [doi: 10.30773/pi.2017.10.2] [Medline: 29614851]
23. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The MINI-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59 Suppl 20:22-33. [Medline: 9881538]
24. Overall JE, Gorham DR. The brief psychiatric rating scale. Psychol Rep 1962 Jun;10(3):799-812. [doi: 10.2466/pr0.1962.10.3.799]
25. HAMILTON M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960 Feb;23(1):56-62. [doi: 10.1136/jnnp.23.1.56] [Medline: 14399272]
26. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med 2001 Jul;33(5):337-343. [doi: 10.3109/07853890109002087] [Medline: 11491192]
27. Sheehan DV. The Anxiety Disease: Charles Scribner’s Sons; 1983:144-153.
28. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]

29. Connor KM, Davidson JRT. Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). Depress Anxiety 2003;18(2):76-82. [doi: 10.1002/da.10113] [Medline: 12964174]
30. Zimet GD, Dahlem NW, Zimet SG, Farley GK. The multidimensional scale of perceived social support. J Pers Assess 1988 Mar;52(1):30-41. [doi: 10.1207/s15327752jpa5201_2]
31. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. 2022 Presented at: NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems; Nov 28 to Dec 9, 2022; New Orleans, Louisiana, USA p. 22199-22213 URL: https://dl.acm.org/doi/10.5555/3600270.3601883 [accessed 2025-12-23] [doi: 10.5555/3600270.3601883]
32. Teng F, Yu Z, Shi Q, Zhang J, Wu C, Luo Y. Atom of thoughts for markov llm test-time scaling. arXiv. Preprint posted online on Feb 17, 2025. [doi: 10.48550/arXiv.2502.12018]
33. Introducing deep research. OpenAI. 2025. URL: https://openai.com/index/introducing-deep-research/ [accessed 2025-12-23]
34. Kwon T, Ong KT, Kang D, et al. Large language models are clinical reasoners: reasoning-aware diagnosis framework with prompt-generated rationales. 2024 Presented at: Proceedings of the AAAI Conference on Artificial Intelligence; Feb 20-27, 2024; Vancouver, Canada p. 18417-18425 URL: https://ojs.aaai.org/index.php/AAAI/article/view/29802 [accessed 2025-12-23] [doi: 10.1609/aaai.v38i16.29802]
35. Wallert J, Boberg J, Kaldo V, et al. Predicting remission after internet-delivered psychotherapy in patients with depression using machine learning and multi-modal data. Transl Psychiatry 2022 Sep 1;12(1):357. [doi: 10.1038/s41398-022-02133-3] [Medline: 36050305]
36. LoParo D, Dunlop BW, Nemeroff CB, Mayberg HS, Craighead WE. Prediction of individual patient outcomes to psychotherapy vs medication for major depression. Npj Ment Health Res 2025 Feb 5;4(1):4. [doi: 10.1038/s44184-025-00119-9] [Medline: 39910171]
37. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. NPJ Digit Med 2023 Oct 20;6(1):195. [doi: 10.1038/s41746-023-00939-z] [Medline: 37864012]
38. Zakka C, Shad R, Chaurasia A, et al. Almanac - retrieval-augmented language models for clinical medicine. NEJM AI 2024 Feb;1(2):AIoa2300068. [doi: 10.1056/aioa2300068] [Medline: 38343631]
39. Borrell-Carrió F, Suchman AL, Epstein RM. The biopsychosocial model 25 years later: principles, practice, and scientific inquiry. Ann Fam Med 2004;2(6):576-582. [doi: 10.1370/afm.245] [Medline: 15576544]
40. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
41. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019 Sep;25(9):1337-1340. [doi: 10.1038/s41591-019-0548-6] [Medline: 31427808]
42. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
43. Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. JAMA 2020 Jan 28;323(4):305-306. [doi: 10.1001/jama.2019.20866] [Medline: 31904799]
44. Obradovich N, Khalsa SS, Khan W, et al. Opportunities and risks of large language models in psychiatry. NPP Digit Psychiatry Neurosci 2024;2(1):8. [doi: 10.1038/s44277-024-00010-z] [Medline: 39554888]
45. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging 2018 Mar;3(3):223-230. [doi: 10.1016/j.bpsc.2017.11.007] [Medline: 29486863]

Abbreviations

AI: artificial intelligence
AoT: atom-of-thoughts
CoT: chain-of-thoughts
HAM-D: Hamilton Depression Rating Scale
LLM: large language model
MAKE BETTER study: MAKE Biomarker Discovery for Enhancing Antidepressant Treatment Effect and Response study
ML: machine learning
NPV: negative predictive value
PPV: positive predictive value
RoD: referencing of deep research
SNRI: serotonin and norepinephrine reuptake inhibitor
SSRI: selective serotonin reuptake inhibitors

Edited by J Torous; submitted 01.Sep.2025; peer-reviewed by E Strobl, Q Wang; revised version received 10.Dec.2025; accepted 11.Dec.2025; published 23.Jan.2026.

Please cite as:
Park JH, Kang HJ, Jeon JH, Kang SG, Kim JW, Kim JM, Lee H
Prediction of 12-Week Remission in Patients With Depressive Disorder Using Reasoning-Based Large Language Models: Model Development and Validation Study
JMIR Ment Health 2026;13:e83352
URL: https://mental.jmir.org/2026/1/e83352
doi:10.2196/83352

Original Paper

Triaging Casual From Critical—Leveraging Machine Learning to Detect Self-Harm and Suicide Risks for Youth on Social Media: Algorithm Development and Validation Study

Sarvech Qadir[1], MSc; Ashwaq Alsoubai[2], PhD; Jinkyung Katie Park[3], PhD; Naima Samreen Ali[4], MSc; Munmun De Choudhury[5], PhD; Pamela Wisniewski[6], PhD

[1]Department of Computer Science, Vanderbilt University, Nashville, TN, United States
[2]Department of Information Systems, King Abdulaziz University, Jeddah, Saudi Arabia
[3]School of Computing, Clemson University, Clemson, SC, United States
[4]School of Information, University of Michigan, Ann Arbor, MI, United States
[5]School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, United States
[6]International Computer Science Institute, ICSI, Berkeley, CA, United States

Corresponding Author:
Sarvech Qadir, MSc
Department of Computer Science
Vanderbilt University
2301 Vanderbilt Place
Nashville, TN
United States
Phone: 1 615 322 7311
Email: sarvech.qadir@vanderbilt.edu

Abstract

Background: This study aims to detect self-harm or suicide (SH-S) ideation language used by youth (aged 13-21 y) in their private Instagram (Meta) conversations. While automated mental health tools have shown promise, there remains a gap in understanding how nuanced youth language around SH-S can be effectively identified.

Objective: Our work aimed to develop interpretable models that go beyond binary classification to recognize the spectrum of SH-S expressions.

Methods: We analyzed a dataset of Instagram private conversations donated by youth. A range of traditional machine learning models (support vector machine, random forest, Naive Bayes, and extreme gradient boosting) and transformer-based architectures (Bidirectional Encoder Representations from Transformers and Distilled Bidirectional Encoder Representations from Transformers) were trained and evaluated. In addition to raw text, we incorporated contextual, psycholinguistic (linguistic injury word count), sentiment (Valence Aware Dictionary and Sentiment Reasoner), and lexical (term frequency–inverse document frequency) features to improve detection accuracy. We further explored how increasing conversational context—from message-level to subconversation level—affected model performance.

Results: Distilled Bidirectional Encoder Representations from Transformers demonstrated a good performance in identifying the presence of SH-S behaviors within individual messages, achieving an accuracy of 99%. However, when tasked with a more fine-grained classification—differentiating among “self” (personal accounts of SH-S), “other” (references to SH-S experiences involving others), and “hyperbole” (sarcastic, humorous, or exaggerated mentions not indicative of genuine risk)—the model’s accuracy declined to 89%. Notably, by expanding the input window to include a broader conversational context, the model’s performance on these granular categories improved to 91%, highlighting the importance of contextual understanding when distinguishing between subtle variations in SH-S discourse.

Conclusions: Our findings underscore the importance of designing SH-S automatic detection systems sensitive to the dynamic language of youth and social media. Contextual and sentiment-aware models improve detection and provide a nuanced understanding of SH-S risk expression. This research lays the foundation for developing inclusive and ethically grounded interventions, while also calling for future work to validate these models across platforms and populations.

(JMIR Ment Health 2026;13:e76051) doi:10.2196/76051

KEYWORDS
suicide/self-harm; machine learning; mental health; youth; natural language processing

Introduction

Background

Youth increasingly turn to social media platforms to express their feelings—even on highly sensitive topics such as self-harm or suicide (SH-S) [1-3]. These platforms provide an outlet where young individuals can seek support and solidarity by connecting with others who share similar struggles [4,5]. The anonymity and expansive reach of these networks allow users to express themselves more freely than in offline contexts; however, such openness can also expose them to potentially harmful content and triggering material that may exacerbate mental health issues [6]. In response, there have been growing efforts to predict SH-S behavior on social media using automated techniques [7,8]. The integration of machine learning (ML) and natural language processing (NLP) into mental health applications shows promising potential for early detection and intervention [9-11]. For instance, artificial intelligence (AI) chatbots implemented in schools to support student mental health have demonstrated how ML-powered technologies can address gaps where human counselors are unavailable [12]. As mental health challenges among youth rise, there is growing interest in using language as a window into psychological well-being. In online settings where traditional support structures may be absent, individuals often express distress through subtle shifts in everyday communication. NLP research in mental health has increasingly focused on identifying language-based signals of psychological distress in everyday communication. For instance, social media studies have shown that individuals experiencing depression or anxiety tend to use more first-person singular pronouns (eg, "I" and "me"), negative emotion words (eg, "sad" and "worthless"), and cognitive processing terms (eg, "think" and "understand") [13]. Other research highlights that temporal focus—such as a shift toward past-tense verbs—and the use of absolutist language (eg, "always," "nothing," and "never") are predictive of emotional dysregulation and suicidal ideation [14]. Individuals who frequently use absolutist terms may be more prone to anxiety and depression [15]. Additionally, research suggests that languages with obligatory future tense marking (eg, "will go" vs "go tomorrow") are associated with lower national suicide rates, possibly by promoting a more future-oriented cognitive framework [16]. Beyond lexical cues, researchers have used syntactic complexity, semantic coherence, and sentiment trajectories to model psychological states across platforms like Reddit, Twitter (subsequently rebranded as X), and Tumblr [17]. However, when applied to the informal and unstructured language of social media, these systems face substantial challenges as youth often use dynamic slang, hyperbolic expressions, and indirect language cues [18-20], which can lead automated models to misinterpret benign or humorous exaggerations as genuine distress signals [21]. The resulting risk of overreaction or underreaction underscores the critical need for context-sensitive approaches that balance timely intervention with the preservation of safe expressive spaces [22-25].

In this work, we explore the effectiveness of combining traditional ML and transformer-based models (eg, Distilled Bidirectional Encoder Representations from Transformers [DistilBERT]), enriched with contextual, psycholinguistic, sentiment, and lexical features. A key innovation lies in our systematic evaluation of contextual window sizes, from single messages to extended subconversations, to determine how different levels of context affect model performance across nuanced SH-S categories: personal disclosures ("self"), references to others' experiences ("other"), and nonserious, exaggerated mentions ("hyperbole"). We move beyond binary classification by systematically evaluating a 3-way schema "self," "other," and "hyperbole," which captures subtle distinctions in SH-S expressions, with implications for more precise triaging. To our knowledge, this is the first study to analyze how varying context modeling strategies impact fine-grained SH-S classification performance using real-world, private youth data. By explicitly modeling the interplay between message content and its conversational context, we aim to develop more accurate, interpretable, and ethically grounded models for SH-S detection. These findings have direct implications for designing safer, context-aware digital interventions that preserve youth agencies while supporting mental health.

Previous Work

Social media platforms are increasingly recognized as pivotal in identifying individuals exhibiting signs of SH-S [26]. Significant strides have been made by ML and social computing scholars in addressing mental health issues like SH-S through automated approaches [27]. Common practices in SH-S detection on social media include advanced ML and NLP techniques, such as Bidirectional Encoder Representations from Transformers (BERT)–based transformer architectures, which capture semantic meanings even when explicit keywords are absent, enabling the identification of at-risk individuals by detecting subtle linguistic markers associated with mental distress [8,27,28]. Despite these advances, such approaches have not been fully adapted to the specific language patterns used by youth on social media [29]. Youth share their experiences with SH-S in diverse ways, such as through hyperbolic expressions (without intent), discussions of others' SH-S experiences, and explicit self-disclosures. A qualitative study by Ali et al [30] revealed that youth disclosures range from hyperbolic expressions (without intent) to discussions of others' SH-S experiences and explicit self-disclosures. Furthermore, previous research has identified distinctions within SH-S, such as differentiating between suicidal ideation, the likelihood of a suicide attempt, and shifts toward suicidal ideation [7]. Most existing studies have employed binary classification methods for SH-S detection, failing to capture the multidimensionality of youth disclosures [31]. Given that online risks exist on a spectrum and can intensify over time, models need to differentiate among varying levels of risk rather than simply categorizing interactions as risky or not risky [32]. While substantial research has provided insights into SH-S prediction on public forums [33,34], there is limited understanding of how such risky discussions unfold in youth private discourses [29]. Our work addresses this multidimensional nature of youth SH-S disclosures in private conversation by performing multiclass classification. Furthermore, previous work identified a key limitation of existing SH-S detection systems is that, while they effectively capture textual content using features such as n-grams, bag-of-words, and deep learning–based word embeddings [35], they often overlook critical contextual factors and user characteristics, such as age and mental health history. This gap can lead to misclassifications that either trigger inappropriate interventions or miss genuine signals of distress [36,37]. To address these challenges, recent work has emphasized a human-centered machine learning approach, which integrates insights from individuals with lived experiences and those providing support to ensure that models are both accurate and interpretable [38]. Incorporating first-person annotations has proven particularly beneficial, as models trained on insider labels tend to outperform those based on third-party annotations [34,39,40]. Furthermore, transparent and explainable AI methods—such as attention-based models and Shapley Additive Explanations [41,42]—are being leveraged to highlight key features driving model decisions, thereby enhancing trust and accountability. Previous studies have shown that linguistic markers, such as absolutist terms, temporal focus, and future-oriented phrasing, are closely linked to mental health outcomes like depression, anxiety, and suicidal ideation, which underscore the value of subtle lexical and syntactic cues in predicting psychological risk [14]. Our work extends this line of research by exploring how such language patterns appear in youth social media conversations. Finally, integrating both linguistic and psychological signals, including sentiment shifts, interaction patterns, and user history [43], is crucial for accurately distinguishing between different forms of SH-S language, such as humorous references, self-disclosures, and discussions of others. By leveraging human-centered machine learning principles, we aim to develop a system that not only detects risk with high accuracy but also respects the agency and lived experiences of individuals expressing distress online.

Goal of This Study and Research Questions

This study aims to develop a context-sensitive framework for detecting and classifying diverse types of SH-S language in youth private communications. Using data from the Instagram Data Donation (IGDD) project [44], where youth (aged 13-21 y) self-labeled their private direct messages as safe or unsafe, we build upon previous annotations of 2019 subconversations from 151 participants [30]. The annotated data are divided into three categories: (1) Self-disclosures of SH-S—explicit personal disclosures of SH-S ideation, (2) SH-S experiences of others—discussions involving SH-S incidents of others, and (3) hyperbolic representations of SH-S—metaphorical or humorous expressions of SH-S–related language.

The study is driven by 2 research questions (RQs). First, what is the best approach to distinguish between different types of SH-S language (humorous, disclosures about others, and self-disclosures) at the message and subconversation levels? Second, what are the contextual and linguistic characteristics that help distinguish between the 3 types of SH-S discourse?

To address these questions, we evaluated several ML models—including transformer-based architectures (BERT and DistilBERT) and classical classifiers (support vector machine [SVM], random forest, Naive Bayes, and extreme gradient boosting [XGBoost])—and enriched them with features such as contextual, psycholinguistic (linguistic injury word count [LIWC]), sentiment (Valence Aware Dictionary and Sentiment Reasoner), and lexical (term frequency–inverse document frequency) indicators. DistilBERT achieved the highest accuracy (99%) in distinguishing SH-S messages. Expanding the context from single messages to subconversations improved model accuracy to 91%. For RQ2, results highlighted that males tended to use hyperbolic SH-S language, females discussed others' experiences, and nonbinary individuals more often shared personal SH-S experiences.

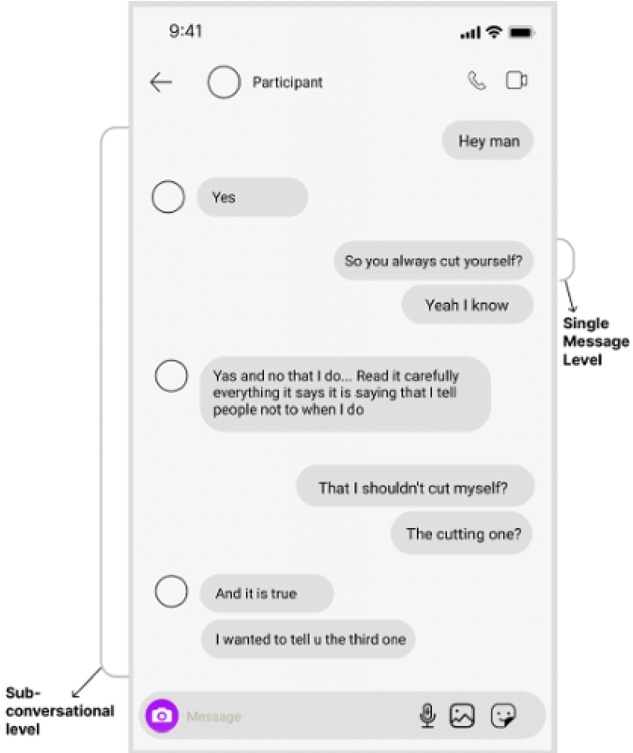
Methods

This section outlines the dataset, preprocessing steps, feature engineering, classification models, and evaluation strategies used to address our research questions.

Dataset

We used a subset of the IGDD project dataset by Razi et al [44], specifically scoped and annotated by Ali et al [30]. In the study by Razi et al [44], researchers collected ecologically valid social media data through the IGDD project to study adolescent online safety. They recruited 195 English-speaking adolescents (aged 13-21 y) in the United States who had active Instagram (Meta) accounts during their teenage years and had experienced at least 2 direct message conversations that made them or someone else feel unsafe or uncomfortable. The dataset includes 32,055 private conversations, providing valuable insights into youth social media interactions and supporting ML models for online risk detection. Ali et al [30] manually annotated the original IGDD dataset for SH-S–related conversations, which resulted in 1224 SH-S–related conversations from 151 youth. The annotators had access to the entire conversation threads, and interrater reliability was calculated (Cohen κ=0.76) during the relevancy coding process to ensure consistency among the annotators in flagging the conversations as involving SH-S language. Interrater reliability was not calculated for the qualitative analyses because the thematic analysis process followed involved an inductive coding process where codes were developed and refined iteratively rather than applied from predetermined categories [45]. Since individual conversations could cover multiple topics, they further segmented them into distinct, topic-specific segments, referred to as subconversations. This process resulted in 2019 subconversations and 35,963 messages from the original 1224 SH-S–related conversations (Figure 1).

Figure 1. Message level and subconversational level.



Annotation Categories

Each subconversation in the SH-S dataset (n=2019) was annotated to capture context-specific interactions and categorized into 3 primary classes.

Self-Disclosures of SH-S

Instances where users explicitly disclose personal SH-S ideation (eg, "I've been cutting again. I don't know how to stop."). This category comprises 1262 (62.5%) subconversations and a total of 10,407 (28.9%) messages.

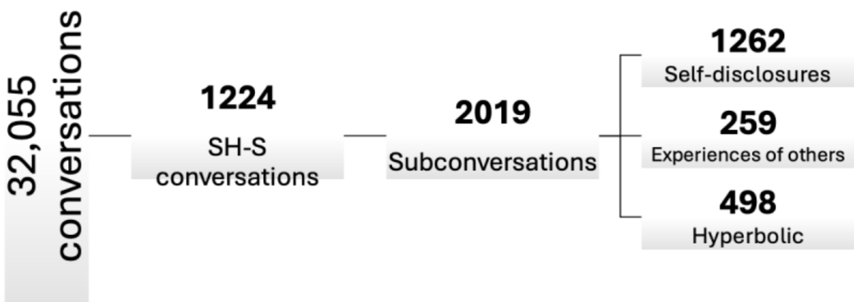
SH-S Experiences of Others

Conversations discussing SH-S incidents involving others (eg, "Did you know she cut themselves?"). This category includes 259 (12.8%) subconversations and a total of 7507 (20.9%) messages.

Hyperbolic Representations of SH-S

Instances where SH-S–related language is used humorously without serious intent (eg, "This homework is killing me!"). This category accounts for 498 (50.2%) subconversations and a total of 18,049 (35.7%) messages (Figure 2).

Figure 2. Overview of dataset distribution. SH-S: self-harm or suicide.



Data Preprocessing

We applied a multistep preprocessing pipeline to clean the dataset while preserving essential linguistic cues.

1. Noise removal: Punctuation, hyperlinks, stop words, non-Latin words, and isolated numeric or single characters were removed. Emojis were converted to text using the Python demoji library to retain emotional cues.
2. Sensitive lexicon development: A lexicon of SH-S and violence-related keywords was constructed based on existing literature [46], guiding the filtering of messages containing SH-S disclosures.
3. Text embedding and filtering: Subconversations were embedded using Sentence-BERT (SBERT), and cosine similarity was computed against the sensitive lexicon. SBERT's proven effectiveness in sentence-level retrieval and paraphrase detection [47] allowed us to identify indirect SH-S references.

4. Data augmentation: BertAug [48] was applied to increase the size and diversity of the training data by generating semantically coherent variations, thereby enhancing model robustness [49]

Table 1 represents the final number of subconversations for each category of SH-S.

Table 1. Number of subconversation instances before and after augmentation.

Category	Original instances, n	Post augmentation, n
Self-disclosures	498	996
SH-S[a] experience of others	259	518
Hyperbolic representations	1262	2592
Total	2019	4106

[a]SH-S: self-harm or suicide.

The final dataset contained 4106 subconversations, offering a more balanced distribution across the 3 categories. To assess the impact of conversational context, we evaluated our best-performing model on the SH-S versus non–SH-S dataset at both the message-level (isolated messages) and the sub-conversational level (15-20 messages of holistic units). Previous research indicates that SH-S language is context dependent [50].

Automatic SH-S Detection Approaches

To address RQ1, we implemented a diverse set of models:

- Transformer-based models: Fine-tuned pretrained BERT [51] and DistilBERT [52] models, leveraging their deep contextual understanding [53] and ability to capture subtle SH-S cues [54].
- Sequential models: Long short-term memory [55], convolutional neural network - bidirectional long short-term memory, and gated recurrent unit [56] architectures were used to model long-term dependencies and temporal patterns in subconversations.
- Classical ML Models: SVM, random forest, Naive Bayes, and XGBoost [57] were trained on engineered features to provide interpretable baselines [50].

Analysis: Feature Engineering

For RQ2, we used a traditional non–end-to-end model (XGBoost, selected for its binary classification performance [57]) using features extracted from the annotated subconversations.

- Contextual features: Demographic data (ie, sex from IGDD survey responses) to explore influences on SH-S discourse.
- Psycholinguistic attributes: LIWC scores [58] covering affective processes (eg, sadness and anger), cognitive mechanisms (eg, causation and certainty), and social or personal concerns, normalized by subconversation length.
- Sentiment analysis: Polarity scores computed using Valence Aware Dictionary and Sentiment Reasoner [59] to capture the emotional tone.
- Lexical features: Term frequency–inverse document frequency [60] scores for the top 1000 terms to measure term importance.
- Additional linguistic scores: Toxicity, politeness, humor, empathy, and hate speech scores extracted using pretrained models from Hugging Face [61].

Model Evaluation

Models were evaluated using standard metrics: accuracy, precision, recall, F_1-score, area under the curve (AUC), and receiver operating characteristic curve. To mitigate overfitting, we used stratified k-fold cross-validation with k=5, which maintains the original class distribution across each fold. This method provided reliable performance estimates, especially in imbalanced datasets [62]. For deep learning models, validation sets were further used to apply early stopping based on validation loss to prevent overfitting [63]. Each model underwent hyperparameter optimization to maximize performance. For classical ML models (SVM, random forest, and XGBoost), hyperparameters (eg, SVM kernel type, XGBoost learning rate, max depth, and number of estimators) were optimized using grid search with cross-validation, which ensures that the hyperparameter evaluation is reliable and generalizable, while grid search exhaustively identifies the best combination of hyperparameters [64]. For deep learning models (long short-term memory, convolutional neural network - bidirectional long short-term memory, and gated recurrent unit), the hyperparameters—such as hidden layer sizes, dropout rates, learning rates, and batch sizes—were optimized using a random search combined with early stopping. For transformer models (BERT and DistilBERT), fine-tuning was performed with learning rate schedulers (eg, linear decay), with hyperparameters like learning rate, batch size, and warm-up steps optimized based on validation loss.

For RQ1, we fine-tuned DistilBERT (distilbert-base-uncased) separately for both message-level and snippet-level classification tasks. The dataset was split into 70% training, 10% validation, and 20% test, using stratified sampling to preserve class distributions across all splits. The test set was completely held out and used solely for final evaluation. All models were trained using the AdamW optimizer with a learning rate of 5e-5, batch size of 16, and a maximum input length of 512 tokens. We trained for up to 10 epochs with early stopping based on validation loss (patience=2) to prevent overfitting. For each classification task, we performed 5-fold cross-validation on the training set and reported average validation results across folds. To reduce overfitting and improve generalization, we applied data augmentation, including back-translation, synonym replacement, and paraphrasing techniques. For each class, we generated augmented samples to approximately double the training set size per class. Each original text was augmented up to 5 times, depending on the number of additional examples needed for that class. The augmenter used its default replacement probability of 0.3, substituting words with semantically appropriate synonyms. For back-translation, we translated text from English to German, and back to English, which added natural lexical and syntactic variation while preserving meaning. We further ensured model stability through a power analysis [65] to confirm that the training data size was sufficient to detect performance differences across model variants with acceptable statistical power. The reported performance metrics reflect results on a completely held-out test set and were consistent across folds and evaluation phases, supporting model generalization and robustness.

Unpacking Results Qualitatively

To gain deeper insights beyond quantitative metrics, we conducted a qualitative reading [35] focused on unpacking model misclassifications, feature importance, and demographic influences on SH-S discourse. First, we examined instances where models failed at the message level but succeeded at the subconversational level in RQ1. Second, unpacking psycholinguistic feature contributions and additional features. We conducted a detailed analysis of the LIWC features [58], evaluating the influence of its 64 categories on model predictions. This involved both quantitative and qualitative steps, such as using Shapley Additive Explanations values, and we identified which LIWC categories most significantly influenced the model's classification decisions. To validate these findings, we manually reviewed subconversations with high activations in key LIWC categories. Furthermore, to understand the significance of each of the additional linguistic features, that is, toxicity [66], politeness, humor, empathy, and hate speech [67], we used ANOVA tests [68] to help understand if there was any significance in prediction. Finally, we conducted demographic analysis (sex and discourse types). To provide additional context on the dataset, we include demographic statistics from the IGDD survey responses, which describe the characteristics of the 151 youth participants whose conversations were analyzed in this study. Most participants identified as female (n=107, 71%), followed by male (n=30, 20%), and nonbinary or preferred not to self-identify (n=14, 9%). For RQ2, we also explored potential demographic influences on SH-S discourse. We conducted a chi-square [69] to examine the relationship between sex and different types of discourse (self-disclosure, SH-S experiences of others, and hyperbolic representations). The test revealed statistically significant differences (P<.05) in discourse types across sex groups.

Ethical Considerations

The secondary analysis of this dataset was reviewed and approved by the Vanderbilt University Institutional Review Board (IRB #222197). The original user study in which the data were collected was approved by the University of Central Florida Institutional Review Board (IRB #00001136). In the original study, informed consent was obtained from all participants; for any participants who were minors, consent or assent procedures were completed in accordance with the approving ethics board requirements. Participants in the IGDD study were compensated with a US$50 Amazon gift card for their time and data contribution. All team members who accessed the data completed required human subjects protections (Collaborative Institutional Training Initiative) and Protection of Minors training before working with the dataset. Moreover, we had a child abuse and imminent risk reporting protocol in place, and any concerning posts were reviewed by the Director of Risk Management and Child Protection to ensure that we met mandated reporting requirements. To protect participant privacy, all data were deidentified during the scoping process. In addition, any quotations included in this manuscript have been edited to remove or mask names, locations, and other potentially personally identifiable details while preserving their substantive meaning.

Results

Overview

In this section, we first present the results of the classifiers that predict SH-S risks at the message level (ie, binary classification). The classifiers were further evaluated with the best-performing end-to-end model at the message and subconversational level (RQ1). Next, we added different features to the best non–end-to-end model (RQ2), and the results were presented, highlighting the unpacking of contextual analysis of our model.

Automated Classification of SH-S in Private Conversations of Youth (RQ1)

Binary Classification (Message Level)

For the classification task to identify SH-S versus non–SH-S messages (binary classification), the DistilBERT (end-to-end learner; accuracy=0.99, precision=0.99, F_1-score=0.99, recall=0.98, and AUC=0.99) model outperformed the other models used across all assessed metrics (Table 2).

Table 2. Performance metrics of various models in binary classification with non–self-harm or suicide data, categorized as end-to-end learners or non–end-to-end learners.

Models	Accuracy	Precision	Recall	F_1-score	AUC[a]	Type
Random forest	0.85	0.88	0.73	0.80	0.83	Non-E2E[b]
SVM[c]	0.89	0.88	0.84	0.86	0.88	Non-E2E
Naive Bayes	0.80	0.87	0.57	0.69	0.91	Non-E2E
XGBoost[d]	0.94	0.91	0.94	0.93	0.98	Non-E2E
CNN[e]	0.88	0.88	0.83	0.85	0.95	E2E
CNN-BiLSTM[f]	0.89	0.85	0.88	0.87	0.90	E2E
LSTM[g]	0.85	0.83	0.79	0.81	0.90	E2E
GRU[h]	0.84	0.82	0.78	0.80	0.92	E2E
BERT[i]	0.97	0.92	0.98	0.95	0.99	E2E
DistilBERT[j]	0.99	0.99	0.98	0.99	0.99	E2E

[a]AUC: area under the curve.
[b]E2E: end-to-end.
[c]SVM: support vector machine.
[d]XGBoost: extreme gradient boosting.
[e]CNN: convolutional neural network.
[f]BiLSTM: bidirectional long short-term memory.
[g]LSTM: long short-term memory.
[h]GRU: gated recurrent unit.
[i]BERT: Bidirectional Encoder Representations from Transformers.
[j]DistilBERT: Distilled Bidirectional Encoder Representations from Transformers.

This high level of accuracy was maintained even when keywords associated with SH-S were removed from the dataset. The results from the keyword removal process using SBERT, which was a methodological approach, also indicated a notable performance with the DistilBERT model, achieving high metrics: accuracy at 0.88, F_1-score at 0.87, precision at 0.80, recall at 0.95, and an AUC of 0.94.

3-Class Classification (Message and Subconversational Level)

Next, we compared the DistilBERT model's performance on a single message and a subconversational level. As seen from Table 3, the inclusion of context at the subconversational level resulted in a marked improvement in the model's performance, with all the accuracy metrics showing an increase.

Table 3. Performance metrics at the message and subconversation levels.

Level	Precision	Recall	F_1-score	Accuracy
Message				0.89
Overall	0.89	0.89	0.89	
Hyperbole	0.90	0.97	0.94	
Other	0.86	0.78	0.82	
Self	0.86	0.73	0.79	
Subconversation				0.91
Overall	0.91	0.91	0.91	
Hyperbole	0.94	0.94	0.94	
Other	0.90	0.95	0.92	
Self	0.83	0.79	0.81	

Our qualitative analysis of classification results confirmed that end-to-end models like DistilBERT, which learn to predict directly from raw text to classification labels, benefited from larger context windows. Specifically, we observed an overall improvement across all performance metrics when moving from message-level to subconversation-level classification. The overall accuracy increased from 0.89 to 0.91, reflecting more accurate predictions with a broader conversational context. This trend was even more pronounced within individual classes. For the "other" class, F_1-score improved from 0.82 to 0.92, and for the "self" class, from 0.79 to 0.81. The model was unable to determine the true intent behind potentially alarming situations when analyzing single messages in isolation. Specifically, the model initially categorized self-disclosure messages indicating imminent risk as hyperbole, likely due to the common use of exaggerated language in casual conversations, particularly among youth. However, responses from conversation partners often provided critical interpretive signals that helped the model correct its classification when additional context was introduced.

For example, in the following subconversation, the model initially classified the message "*And want to kill myself*" as hyperbole risk. Yet, when the model processed the entire subconversation, it identified contextual cues such as "*No don't kill yourself*" indicating that the situation was serious rather than exaggerated and highlighting an ongoing struggle and the urgency of the distress being expressed.

> *P: And want to kill myself*
>
> *O: I can't believe you have to deal with this everyday*
>
> *O: No don't kill yourself, It's not a good idea*
>
> *P: I'm killing myself*

For 2-way conversations, "P" refers to the primary participant, whose conversation is being analyzed. "O" denotes the other individual involved in the exchange. For group conversations, we appended numbers "O1," "O2," "O3," and so on, to denote the other individuals participating in the group conversation. Conversely, the opposite scenario occurred when the model misclassified messages as immediate "self" risk. Upon further examination of the context, it became clear that the youth was either exaggerating their emotions or discussing the struggles of someone else. This was likely because the messages appeared to directly address an individual who may be struggling with suicidal thoughts or because they contained words like "cut" and "I," which are often associated with personal distress and self-harm. The pronoun "I" typically signals self-referential language, which, when combined with distress-related terms, may lead the model to interpret the message as expressing suicidal intent. However, additional context, such as the presence of laughter, such as "HAHAHAHAHA," provided crucial disambiguation, shifting the interpretation toward humor or hyperbole. In the following subconversation, the initial classification of self-risk was corrected when the broader exchange revealed that the discussion was playful rather than an actual self-harm disclosure:

> *O: Can you tell me The frick, Did you just cut, HAHAHAHAHA*
>
> *P: It's my middle*
>
> *O: I DON'T CARE, WHYD YOU CUT, NOO, Are you kidding me. COME ON,*
>
> *WIPE IT UP, And I'll tell you, And take a picture of it cleaned*
>
> *O:Put a bandaid on your cut*
>
> *P: I didn't cut myself, It was food I KNEW IT. I knew I used a different finger to draw red*

In addition, when additional context from the broader subconversation was added, it became evident that the participants were not discussing their own distress but rather engaging in third-person discussions about distressing events. For example, in the following subconversation, the message was initially classified as "self," but with added context, it became clear that the participants were reacting to the news of someone else's passing.

> *P: How are you feeling*
>
> *O: Still in shock but, life goes on. Gotta keep moving forward*
>
> *P: And you should know, suicide is not the answer, think about the good things in life*
>
> *O: Yeah, you have to try and find peace*
>
> *P: Yeah I couldn't believe it. This is a sad news*
>
> *O: May he rest in peace. I knew this guy*

The improved performance at the subconversational level demonstrates the benefit of incorporating context, providing richer semantic information than isolated messages.

Table 4 below presents a qualitative breakdown of frequent misclassifications. The most common confusion occurred between self-disclosure and hyperbolic expression, particularly when tone was ambiguous or sarcasm was present. Other frequent errors stemmed from third-person framing, generalizations, or lack of conversational context, underscoring the limitations of message-level models.

Table 4. Examples of misclassifications between message and subconversational levels.

Misclassified type	Misclassified cases (actual), n (%)	Example message	Example subconversation	Likely cause
Other to self	5 (6.02)	Who repeatedly threatened suicide and harassed her if she didn't spend every single free moment talking to him?	Have you ever seen [person'] video on her ex boyfriend? Who repeatedly threatened suicide and harassed her if she didn't spend every single free moment talking to him? She met him in person and thought she was safe because he was a scrawny nerd and thought "I can totally fight him if he forces himself on me." She couldn't defend herself when she really needed it because she had been too mentally worn down by her boyfriend to resist and froze."	Third-person narrative was judged as suicide due to lack of context
Hyperbole to self	15 (18.07)	guys i'm being stupid bc ive continued to snap thomas and like minutes ago he snapped me "you cute :)" and i literally wanna kms i'm disappointed in myself also to add insult to injury he was in his boy scount uniform	guys i'm being stupid bc ive continued to snap thomas and like minutes ago he snapped me "you cute :)" and i literally wanna kms i'm disappointed in myself also to add insult to injury he was in his boy scount ahhhh-damnnn you got a boy scout noooo he's not my lol you right he's your boy	Sarcasm, dramatic and violent tone interpreted literally as personal disclosure
Hyperbole to other	2 (2.41)	She is like,, talking to everyone in the groupchat and just like about to die lol	"I do need to get over this disgusting possessiveness but it could be good! not weird! And I'm not gonna private lol that's weirddddd Not flirty! She is just talking normally but my possessiveness is getting BAD how was your orientation tho!!! she prolly just has a flirty personality tho which sucks i think you should talk to her in private message!!! lmao Well I have job orientation so I couldn't join lmoa They voice chatted w/o me too : She is like,, talking to everyone in the groupchat and just like about to die lol My jealousy/possessiveness is getting REALLY bad lol"	Ambiguous intent; informal tone predicted as someone else's disclosure
Other to hyperbole	15 (18.07)	Wow but u up future [me] to kill myself	P: ['Tomorrow! Guess I ' ll keep him. .. O: lol I cannotttttttt P: Exactly O: ooooh dad Meee P: Wow but u up future [me] to kill myself P: Omg describes me still to dry out feet Ok [person]]	Figurative expressions misread as hyperbole
Self to other	7 (8.43)	my friend sent me a suicide message	P: and she wasn't responding for at least minutes to an hour P: my friend sent me a suicide message- O: that's a lot of chat to read O: lemme read the chat O: wait? what happened? P: she texted me, luckily she's still here but she barely has the strength to keep moving O: I had to get the sword out just to use it to reach a pair of scissors I accidentally got stuck somewhere-	Third-person framing; lack of first-person indicators in the message level for prediction of Self category. Expanded context in sub-conversational level
Self to hyperbole	39 (46.99)	I did and I'm gonna continue it bc it won't let me die this week next week I'll continue to be dead	P: "I'd rethink my life choices My childhood Wellfuq private O: Ill join P: I did and I'm gonna continue it bc it won't let me die this week next week I'll continue to be dead O: I CRI ERYTIME FAM I THOT U ENDED URSRLF P: YESSS P: I literally thought i was going to get shit done but then i realized we've been blessed with the internet"	Overlapping emotional expression and slang

Contextual and Psycholinguistic Characteristics of Youth SH-S Conversation (RQ2): Psycholinguistic Characteristics of Youth SH-S Conversation

We further unpacked the linguistic features and presented the results from a qualitative analysis evaluating how the psycholinguistic characteristics were associated with classification results for each category (self, other, and hyperbole). We completed this analysis at the subconversation level rather than the message level due to more context, as confirmed by RQ1 results. Our results show that psycholinguistic words in each subconversation helped predict each category. For example, the words associated with tentativeness ("tentat") showed high importance for prediction in the "self" category. This suggested that individuals speaking from personal experience expressed uncertainty or hesitation, which could reflect the complexity of their feelings or situations. This was seen in the use of words like "maybe," "or," "anyone," "if," and "anything." For example, in the following conversation, tentative words helped with the prediction of a participant reminiscing about something from their past:

> *P: i was all like "oh man what if this all gets so bad that i like self harm or something i'm going to cry" because i wanted attention and then fifth grade came and i was just fine for no reason*
>
> *O: haha so i remember i was like "if i don't get this many likes in this many days i'm leaving"*
>
> *P: i remember posting this thing about how i was all "depressed"*

Furthermore, words associated with space, such as "spatial" and "time" language, were prominent for the prediction of the "hyperbole" category, which implied a discussion about SH-S that involves words related to physical spaces or distances. This may be relevant in conversations where individuals are talking about other people or situations removed from themselves. For example, in the following subconversation, spatial terminology played a crucial role in amplifying the description of a strenuous physical routine. The words underlining physical locations and movements, such as "in the heat," "up and down hills," "on our track," "on the hot hard ground," and "on the track," constructed a scenario rife with exertion. The hyperbolic essence is further enhanced by the repetitive mention of these terms, intensifying the narrative. The speaker used this spatial language to dramatically overstate and exaggerate, typical of hyperbolic discourses. For instance, phrases like "weird ass suicides on the track" and the exaggerated claim of being ready to "pass out" further illustrated the intensity of the workout for dramatic effect.

O1: nahhh we be running with full pads in the heat up and down hills and the sun shine directly on our track

> *O2: in the afternoon we run a mile do drills where we gotta lay down on the hot hard ground*
>
> *O3: we do bleachers we do laps if our group came in last we do mad push ups crutches etc.. we do weird ass suicides on the track.*

In addition, language focused on family-related words ("home") and social processes ("social") aided in the prediction of the "other" category, which implied that discussions about others' situations in a home or family context or interactions when discussing others' experiences with SH-S. For instance, the following conversation talks about how social and family keywords, that is, "roommate," "person," "person1," "names," helped in prediction to models when talking about the experiences of others.

> *P: The one that never showers has a new kid roommate and everyone in the suite beside her is new. Besides some that had accidents the rest r ppl that no one wanted to room with.*
>
> *O: Most returners with new kid roommates? Watch out.*
>
> *P: Yeahhhhhh. And it makes [person] hecka sad and he's been staying in my room with me cause it's too much for him.*
>
> *O: Cause [person] has him and he's dying cause [person1] won't talk about anything except suicide and gays.*
>
> *P: At least his room is clean and he showers better than some annoying ppl.*

As such, the psycholinguistic features aided the model in prediction based on the higher frequency of specific features in individual categories.

Identifying Key Differentiators in SH-S Conversations

To evaluate the role of additional linguistic features in distinguishing between categories (hyperbole or humor, other, and self), a 1-way ANOVA test was conducted. Table 5 presents the results of the ANOVA analysis.

Table 5. ANOVA results for distinguishing conversational categories. Statistically significant features (P<.05) indicate meaningful variations across categories.

Feature	F test (df)	P value
EmpathyScore	369.22 (2, 2141)	<.001
Toxicity	144.63 (2, 2141)	<.001
PolitenessScore	22.64 (2, 2141)	<.001
HumorScore	18.74 (2, 2141)	<.001
HateSpeechScore	1.14 (2, 2141)	.32 (ns[a])

[a]ns: not significant.

Among the features analyzed, EmpathyScore (*P*<.001) was the strongest differentiator, particularly in *self* messages. This suggests that self-harm disclosures often contain highly empathetic language, either from the individual expressing distress or from responders offering support.

> *P: It's not very good please don't get mad. It's how I calm myself. It helpss.*
>
> *O: Oh, I am not mad. What matters to me right now is that you're feeling better...Though, is it okay if I request that we attempt to find a safer method in the future? There are endless ways, we can find one!*

Here, the self message contains hesitation and a plea for nonjudgment, while the response demonstrates high empathy by validating the user's feelings before gently encouraging an alternative. These findings reinforce the importance of empathy detection in SH-S identification—messages with high "EmpathyScore" often signal distress and should be carefully analyzed rather than dismissed as neutral or unimportant. Additionally, toxicity (*P*<.01) was more prevalent in *hyperbole or humor* messages, indicating that sarcasm and exaggeration often introduce aggressive or negative wording. However, this poses a challenge for SH-S detection, as some genuine disclosures may also be flagged as toxic despite containing cries for help.

> *P: I sent this to her and said you included [person] before meeee Fucking kill me, LOL legit and we will always be novices in their eyes even when we are not.*

Here, the phrase "*Fucking kill me*" demonstrates self-harm references with violent references increasing the overall toxicity of the subconversation. Interestingly, toxicity and EmpathyScore were positively correlated (*r*=0.41), highlighting a key challenge identified in content moderation systems; messages flagged as toxic also contain empathy. This indicates that toxicity does not always equate to harmful intent—rather, some individuals express distress in a way that appears toxic but is a plea for help.

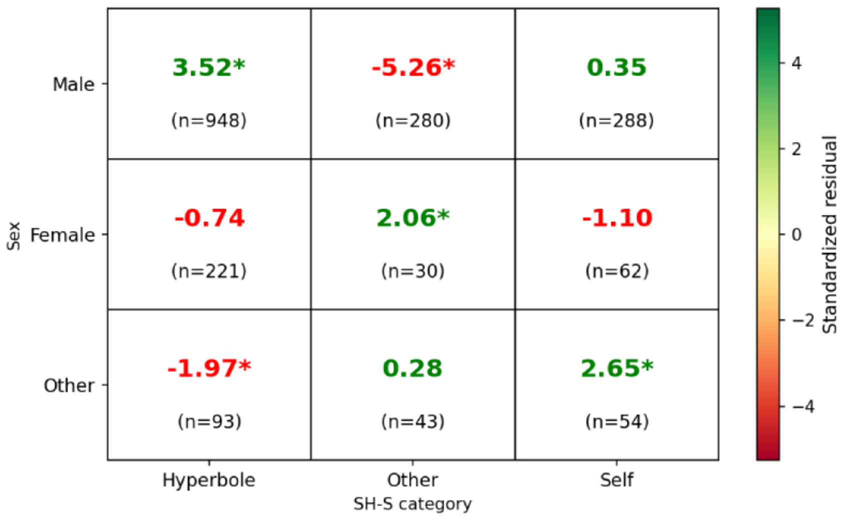
> *P: he would call me a fat whore and a bitch, and would tell me to kill myself, and i stayed with him for six months and cried everyday.*

In this message, the high toxicity score is due to the offensive language describing verbal abuse, yet the high empathy score reflects the emotional weight of the experience being shared. If an AI system were to automatically filter high-toxicity messages as humor, it might wrongly remove important self-harm disclosures, preventing individuals from receiving support. Therefore, there is a need for context-aware, explainable moderation rather than strict toxicity-based filtering.

Contextual Characteristics of Youth SH-S Conversation

Next, we unpacked the contextual factors and presented the results from the relationship between "sex" feature and discourse categories ("hyperbole," "other," and "self,") as captured in Figure 3, highlighting noteworthy patterns in the communication styles among different sexes using a chi-square test. In this analysis, standardized residuals with an absolute value greater than 1.96 (indicated by an asterisk) denote a significant deviation from expected frequencies (*P*<.05).

Figure 3. Residuals: sex versus category. The magnitude of the standardized residual (eg, above 2 or below –2) indicates the strength of this deviation (|z|>1.96; *P*<.05), denoted with (*) with the value. SH-S: self-harm or suicide.



To examine sex-specific patterns in SH-S discourse, we conducted a chi-squared test of independence between sex identity (female, male, and other) and SH-S discourse categories (hyperbole, other, and self). The analysis yielded a statistically significant association (χ^2_4=57.16; n=2019; *P*<.001), indicating that the distribution of SH-S expression types significantly varied by sex. Standardized residuals revealed notable patterns in language use across sex groups (Figure 3). Male participants were substantially more likely than expected to use hyperbolic SH-S language (residual=3.52*), reflecting a frequent use of exaggerated or nonliteral references to SH-S. At the same time, they were significantly less likely to discuss the SH-S experiences of others (residual=–5.26*), suggesting that third-person narratives were underrepresented in male-authored conversations. Participants identifying as nonbinary exhibited a distinct linguistic profile. They were less likely to use hyperbole (residual=–1.97) and more likely to disclose personal SH-S experiences (residual=2.65*), reflecting a stronger inclination toward emotionally vulnerable and direct expression of mental health challenges. Female participants, on the other hand, were more likely to reference others' SH-S experiences (residual=2.06*), indicating a tendency to discuss SH-S in the context of third parties, such as in gossiping situations.

Discussion

Principal Findings

Prior studies have typically focused on binary classification of SH-S or used public social media datasets. In contrast, our work introduces a granular classification of SH-S disclosures. We further examine how varying levels of conversational context influence model performance and incorporate psycholinguistic and contextual signals into transformer-based models to improve interpretability. Our study highlights the effectiveness of using transformer-based models, particularly DistilBERT, for detecting SH-S ideation in youth's private Instagram conversations. DistilBERT achieved 99% accuracy in binary SH-S classification and 91% accuracy in multiclass categorization when extended to subconversational context. Incorporating features, such as sentiment, psycholinguistic cues, and contextual windows, significantly improved model performance. We also uncovered meaningful sex-based patterns in SH-S language: males were more likely to use hyperbolic expressions, females tended to discuss others' experiences, and nonbinary individuals predominantly shared personal disclosures.

The Importance of Context in Detecting SH-S

The classification findings (RQ1) highlight the importance of using our context-sensitive approach, which addresses the limitations of traditional binary classification and the reliance on single messages that constrained the model's ability to capture the nuanced meanings and intentions of SH-S discourses. This approach enabled us to develop an automatic detection model that distinguished between casual expressions and critical SH-S situations, achieving an accuracy of 91%. Although this level of accuracy is noteworthy, we advocate for a paradigm shift in future research. Rather than focusing solely on improving model accuracy or optimizing algorithms [70], greater emphasis should be placed on identifying and understanding the nuanced variations within SH-S–related content, as well as determining the optimal conversational context necessary to inform and enhance the detection capabilities. By incorporating these dimensions into detection models, researchers could differentiate low-risk expressions from those that signal imminent danger and reduce both false negatives, where serious risk may be overlooked, and false positives, where benign expressions are mistakenly flagged, as we showed in our error analysis. Recognizing these distinctions supports more targeted and ethically responsible interventions, ensuring that individuals in genuine need receive appropriate care while minimizing unnecessary responses. Future models could advance this framework by adopting more refined annotation schemes, as previous work has identified various proxies indicative of mental health challenges online. For example, online harassment has been shown to precede declines in mental health [71], and groups of teens who engaged in self-harm offline were found to participate in high-risk sexual conversations with strangers online [72]. These findings, along with the granular approach used in this study, underscore the complexity of identifying SH-S cases in online conversations, far beyond what a simple binary classification can capture. In addition, we recommend future research to leverage advanced techniques, such as reinforcement learning [73], to dynamically identify and integrate the most informative conversational context. Such efforts will advance the field toward a more nuanced, context-aware, and human-centered approach to SH-S detection across both clinical and online environments.

Contextual and Linguistic Implications for Automated SH-S Detection

The prominence of tentative words (eg, "maybe," "if," "anyone," and "anything") in SH-S ideation disclosures suggests that individuals may exhibit uncertainty or hesitation when expressing their personal experiences [74]. This observation aligns with previous psychological research, which indicates that ambivalence is common among individuals contemplating SH-S, reflecting an internal struggle or emotional distress [75]. Consequently, AI models for SH-S detection must be designed to contextualize such uncertainty rather than dismiss it outright. Flagging potential risk cases based solely on tentative language may lead to overlooking individuals in distress; therefore, future work should consider incorporating longitudinal analyses to determine whether repeated use of tentative expressions correlates with escalating distress over time [76]. Furthermore, the model's reliance on spatial references (eg, "on the track" and "up and down hills") and temporal references (eg, "in the heat") for classifying hyperbolic language highlights the role of vivid, metaphorical descriptions in exaggerated narratives. In social media conversations, hyperbole is frequently used as a form of dark humor, exaggeration, or emphasis rather than as an indication of genuine self-harm intent [30]. Previous studies have demonstrated that language not only conveys information but also shapes cognitive processes; for example, spatial and temporal metaphors influence how individuals conceptualize and articulate their experiences [77]. Given the prevalence of such hyperbolic expressions among youth, detection systems must adapt to evolving linguistic trends. The integration of sentiment analysis and contextual embeddings is recommended to enhance classification accuracy by distinguishing between distress-driven hyperbole and casual or humorous exaggeration. The frequent mention of social and familial words (eg, "roommate," "person," and "home") suggests that discussions regarding others' SH-S experiences are often framed within interpersonal and domestic contexts. This linguistic pattern reflects how individuals process and externalize their concerns by embedding them in familiar social environments [78]. Research indicates that discussing traumatic events, such as another's SH-S, can serve as a coping mechanism, helping individuals process their emotions through interpersonal narratives [79]. For automated detection systems, it is critical to differentiate between self-reports and third-party observations. While discussions about others may not signal personal risk, they can indicate a user's concern for someone at risk [80]. Future research should explore networked conversational analysis to identify patterns in how information about SH-S is shared, potentially guiding the design of interventions that support indirect reporting in real-time crisis situations.

Sex-Based Patterns in SH-S Detection

Our analysis reveals distinct sex-based patterns in SH-S communications. Specifically, males in our dataset more frequently used hyperbolic or exaggerated language, females tended to discuss SH-S in the context of others' experiences, and nonbinary youth were more likely to share personal self-disclosures. One potential explanation is that these differences mirror varying psychological needs and coping mechanisms, where males are using violent language [81]. For example, males may use humor or hyperbole as a form of emotional regulation or deflection [82], while females may rely more on social or communal support, thus centering their discussions on peers, friends, or family members [83]. Meanwhile, nonbinary youth could find direct self-disclosure to be a more authentic way to articulate distress, potentially reflecting lived experiences tied to identity-related stressors [84]. Individual communication styles are shaped by intersecting factors, such as cultural background, social context, and personal history, which means no single sex group is homogeneous in its expression of SH-S [85]. Additionally, differences in disclosure may stem from how comfortable individuals feel discussing mental health in private online spaces, rather than from any inherent sex-based communication pattern [86]. Future research should therefore investigate not only what different sex groups share online but also the broader context—how, when, and why they choose to share. An intersectional approach that considers overlapping identities (eg, race, ethnicity, and sexual orientation) may offer deeper insights and help prevent models from misclassifying or overlooking at-risk behaviors. Furthermore, detection systems need to account for these diverse linguistic and psychological signals without reinforcing biases. By doing so, we can develop more inclusive, accurate, and ethically grounded tools for identifying and responding to SH-S discourse across varying demographic groups.

Toward Clinical and Educational Applications

While our study primarily focuses on advancing SH-S detection models through nuanced classification and context-aware techniques, our findings lay important groundwork for translation into real-world clinical and educational settings. One of the major challenges for automated risk detection systems is the high rate of false positives [87], which can overburden already stretched mental health infrastructures and erode trust among stakeholders [88]. Our approach can help address this concern by enabling triaged classification systems (ie, differentiating low-risk, hyperbolic expressions from high-risk, and self-disclosure content) that prioritize urgency and minimize unnecessary escalations. For instance, high-confidence "self" messages may be routed for review by school counselors or clinical professionals, while "other" or "hyperbole" cases can prompt peer-based support, reflective check-ins, or educational messaging. These differentiated outputs could be integrated into existing digital mental health infrastructures such as Crisis Text Line [89], Kognito simulations [90], or school-based early warning systems to assist professionals in identifying emerging risks. Rather than triggering blanket alerts, the model output can inform tiered intervention protocols that optimize limited human resources and reduce unnecessary escalations [91]. Importantly, such integration should not replace human support but rather augment it by flagging concerning patterns at scale while preserving user autonomy and privacy. Furthermore, schools are well-positioned to implement suicide prevention strategies [92] but often rely on direct disclosures, missing the indirect language youth use online [93]. Our work suggests that expanding these systems to include models trained on authentic youth expressions, including hyperbole and third-person narratives, could reduce false negatives and better capture subtle signs of distress. Beyond detection, insights from our findings have important implications for mental health education. In school-based programs, educators can incorporate reflective writing or discussion-based activities using examples of SH-S–related messages to help students explore how emotions are expressed online. For instance, educators can teach youth how to express emotional distress more descriptively to reduce misinterpretation and unintentional triggering. In addition, our findings highlight the need to equip youth with the skills to recognize when peers might be signaling distress, directly or indirectly. These educational opportunities together can help youth not only to seek help when needed but also to support one another in navigating emotionally charged conversations in safe and constructive ways.

Limitations and Future Work

A primary limitation of our study is the reliance on human annotations from Ali et al [30], where the instances analyzed were not flagged by the target participants themselves but were instead annotated by researchers. Additionally, the challenges of obtaining datasets in our domain may limit the generalizability of our findings to other platforms. Another limitation is that, due to the sensitive nature of the data, the original researchers have only shared it on a limited basis with trusted collaborators. This restricted access means that the dataset cannot be made publicly available, thus limiting opportunities for independent validation and replication of our findings. Another limitation identified in our error analysis was that existing models, like DistilBERT, do not explicitly account for slang and may struggle with the informal language used by youth. While some slang and code words, such as "kms" or "dies," were captured due to their frequency in the dataset and contextual embeddings, less common or emerging terms may still be missed. In addition, our model was trained exclusively on English language data from US-based adolescents, which may limit its applicability to non-English speaking or culturally distinct youth populations.

Future studies should apply our approach to other youth datasets where the data is flagged either by the participants themselves or by clinicians. This would help ensure that the models are responsive to the subjective experiences of individuals at risk. Secondly, incorporating insights from clinical experts in reviewing flagged data could provide a more comprehensive understanding of risk levels and improve the classifiers' accuracy. Further work should focus on enhancing ML models by integrating slang and other colloquial expressions to improve their accuracy in detecting SH-S language. Deeper attention should be paid to the detection of evolving code word conversations used by youth to talk about distress in indirect ways. The use of mutually exclusive labels ("self," "other," and "hyperbole") may oversimplify complex expressions that span multiple categories. Future work could explore more flexible annotation schemes—such as multilabel classification or probabilistic tagging—to better capture the fluidity of online discourse. Additionally, incorporating human-in-the-loop testing or clinician usability assessments would offer valuable insight into deployment challenges and model effectiveness in real-world settings. Finally, researchers should work on assessing the timeliness of detection, ensuring that the models classify risks accurately in a time-sensitive manner. Future extensions should also consider applying this work to other languages (eg, Spanish and Chinese) and regions (eg, Asia and Europe) to explore how cultural semantics shape youth expressions of self-harm, sarcasm, and distress.

Conclusion

This study demonstrates the potential of ML models—particularly transformer-based architectures like DistilBERT—to accurately detect SH-S ideation within youth's private social media conversations. By moving beyond binary classification and incorporating contextual, psycholinguistic, sentiment, and lexical features, our approach captures the nuanced spectrum of SH-S expressions, from hyperbole to personal disclosures. Importantly, we find that expanding the context window to subconversations significantly improves classification accuracy, underscoring the critical role of conversational context in understanding youth mental health language. Furthermore, sex-specific patterns in SH-S expression highlight the need for inclusive models that account for diverse linguistic behaviors. As digital platforms become central to youth communication, our findings emphasize the importance of context-aware, ethically designed interventions that can support timely and sensitive mental health responses.

Funding

Researchers who co-authored this work were supported by the US National Science Foundation under grants #IIS-2550812, #TI-2550746, and #CNS-2550834, and by the William T. Grant Foundation grant #187941. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the research sponsors.

Conflicts of Interest

None declared.

References

1. Shensa A, Sidani JE, Escobar-Viera CG, Switzer GE, Primack BA, Choukas-Bradley S. Emotional support from social media and face-to-face relationships: Associations with depression risk among young adults. J Affect Disord 2020;260:38-44 [FREE Full text] [doi: 10.1016/j.jad.2019.08.092] [Medline: 31493637]
2. Best P, Manktelow R, Taylor BJ. Social work and social media: online help-seeking and the mental well-being of adolescent males. Br J Soc Work 2014;46(1):257-276. [doi: 10.1093/bjsw/bcu130]
3. Pretorius C, Chambers D, Cowan B, Coyle D. Young people seeking help online for mental health: cross-sectional survey study. JMIR Ment Health 2019;6(8):e13524 [FREE Full text] [doi: 10.2196/13524] [Medline: 31452519]
4. Kim T, Hong H. Understanding university students' experiences, perceptions, and attitudes toward peers displaying mental health-related problems on social networking sites: online survey and interview study. JMIR Ment Health 2021;8(10):e23465 [FREE Full text] [doi: 10.2196/23465] [Medline: 34609315]
5. Marchant A, Hawton K, Stewart A, Montgomery P, Singaravelu V, Lloyd K, et al. A systematic review of the relationship between internet use, self-harm and suicidal behaviour in young people: The good, the bad and the unknown. PLoS One 2017;12(8):e0181722 [FREE Full text] [doi: 10.1371/journal.pone.0181722] [Medline: 28813437]
6. Naz A, Naureen A, Kiran T, Husain MO, Minhas A, Razzaque B, et al. Exploring lived experiences of adolescents presenting with self-harm and their views about suicide prevention strategies: a qualitative approach. Int J Environ Res Public Health 2021;18(9):4694 [FREE Full text] [doi: 10.3390/ijerph18094694] [Medline: 33924930]
7. Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digit Med 2020;3:43 [FREE Full text] [doi: 10.1038/s41746-020-0233-7] [Medline: 32219184]
8. Ghanadian H, Nejadgholi I, Osman HA. Socially aware synthetic data generation for suicidal ideation detection using large language models. IEEE Access 2024;12:14350-14363. [doi: 10.1109/access.2024.3358206]
9. Kannan KD, Jagatheesaperumal SK, Kandala RNVPS, Lotfaliany M, Alizadehsanid R, Mohebbi M. Advancements in machine learning and deep learning for early detection and management of mental health disorder. arXiv:2412.06147 2024. [doi: 10.48550/arXiv.2412.06147]
10. Elbattah M, Arnaud E, Gignon M, Dequen G. The role of text analytics in healthcare: a review of recent developments and applications. Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021) 2021;5:825-832. [doi: 10.5220/0010414508250832]
11. Hao T, Huang Z, Liang L, Weng H, Tang B. Health natural language processing: methodology development and applications. JMIR Med Inform 2021;9(10):e23898 [FREE Full text] [doi: 10.2196/23898] [Medline: 34673533]

12. W. S. Journal. When there's no school counselor, there's a bot. Wall Str. J. 2025. URL: https://www.wsj.com/tech/ai/student-mental-health-ai-chat-bots-school-4eb1ba55?utm_source=chatgpt.com [accessed 2025-10-27]
13. Wankmüller S. A comparison of approaches for imbalanced classification problems in the context of retrieving relevant documents for an analysis. J Comput Soc Sci 2023;6(1):91-163 [FREE Full text] [doi: 10.1007/s42001-022-00191-7] [Medline: 36568019]
14. Al-Mosaiwi M, Johnstone T. In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clin Psychol Sci 2018;6(4):529-542 [FREE Full text] [doi: 10.1177/2167702617747074] [Medline: 30886766]
15. Lao C, Lane J, Suominen H. Analyzing suicide risk from linguistic features in social media: evaluation study. JMIR Form Res 2022;6(8):e35563 [FREE Full text] [doi: 10.2196/35563] [Medline: 36040781]
16. Lien D, Zhang S. Words matter life: The effect of language on suicide behavior. Journal of Behavioral and Experimental Economics 2020;86:101536. [doi: 10.1016/j.socec.2020.101536]
17. Greaves MM. A corpus linguistic analysis of public reddit and tumblr blog posts on non-suicidal self-injury. Graduate Thesis Or Dissertation 2018. [doi: 10.31234/osf.io/k4qt3]
18. Li X, Chen F, Ma L. Exploring the potential of artificial intelligence in adolescent suicide prevention: current applications, challenges, and future directions. Psychiatry 2024;87(1):7-20. [doi: 10.1080/00332747.2023.2291945] [Medline: 38227496]
19. Nuñez-Rola C, Ruta-Canayong NJ. Social media influences to teenagers. Int. J. Res. Sci. Manag 2019;6:38-48 [FREE Full text]
20. McCulloch G. Because internet: understanding the new rules of language. Revista de Estudios del Discurso Digital 2021(4):146-156. [doi: 10.24197/redd.4.2021.146-156]
21. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Communication 2015;71:10-49. [doi: 10.1016/j.specom.2015.03.004]
22. Lavis A, Winter R. #Online harms or benefits? An ethnographic analysis of the positives and negatives of peer-support around self-harm on social media. J Child Psychol Psychiatry 2020;61(8):842-854. [doi: 10.1111/jcpp.13245] [Medline: 32459004]
23. Choukas-Bradley S, Roberts SR, Maheux AJ, Nesi J. The perfect storm: a developmental-sociocultural framework for the role of social media in adolescent girls' body image concerns and mental health. Clin Child Fam Psychol Rev 2022;25(4):681-701 [FREE Full text] [doi: 10.1007/s10567-022-00404-5] [Medline: 35841501]
24. Keller MC, Nesse RM. The evolutionary significance of depressive symptoms: different adverse situations lead to different depressive symptom patterns. J Pers Soc Psychol 2006;91(2):316-330. [doi: 10.1037/0022-3514.91.2.316] [Medline: 16881767]
25. Sapiro B, Ward A. Marginalized youth, mental health, and connection with others: a review of the literature. Child Adolesc Soc Work J 2019;37(4):343-357. [doi: 10.1007/s10560-019-00628-5]
26. Memon A, Sharma S, Mohite S, Jain S. The role of online social networking on deliberate self-harm and suicidality in adolescents: a systematized review of literature. Indian J Psychiatry 2018;60(4):384-392 [FREE Full text] [doi: 10.4103/psychiatry.IndianJPsychiatry_414_17] [Medline: 30581202]
27. Zeberga K, Attique M, Shah B, Ali F, Jembre YZ, Chung T. A novel text mining approach for mental health prediction using Bi-LSTM and BERT model. Comput Intell Neurosci 2022;2022:7893775 [FREE Full text] [doi: 10.1155/2022/7893775] [Medline: 35281185]
28. Crema C, Attardi G, Sartiano D, Redolfi A. Natural language processing in clinical neuroscience and psychiatry: a review. Front Psychiatry 2022;13:946387 [FREE Full text] [doi: 10.3389/fpsyt.2022.946387] [Medline: 36186874]
29. Forte A, Sarli G, Polidori L, Lester D, Pompili M. The role of new technologies to prevent suicide in adolescence: a systematic review of the literature. Medicina (Kaunas) 2021;57(2):109 [FREE Full text] [doi: 10.3390/medicina57020109] [Medline: 33530342]
30. Ali NS, Qadir S, Alsoubai A, De Choudhury M, Razi A, Wisniewski PJ. "I'm gonna KMS": from imminent risk to youth joking about suicide and self-harm via social media. 2024 Presented at: CHI '24: CHI Conference on Human Factors in Computing Systems; 2024 May 11 - 16; Honolulu HI USA p. 1-18. [doi: 10.1145/3613904.3642489]
31. Haque R, Islam N, Islam M, Ahsan MM. A comparative analysis on suicidal ideation detection using nlp, machine, and deep learning. Technologies 2022;10(3):57. [doi: 10.3390/technologies10030057]
32. Savoia E, Harriman NW, Su M, Cote T, Shortland N. Adolescents' exposure to online risks: gender disparities and vulnerabilities related to online behaviors. Int J Environ Res Public Health 2021;18(11):5786 [FREE Full text] [doi: 10.3390/ijerph18115786] [Medline: 34072253]
33. Brennan C, Saraiva S, Mitchell E, Melia R, Campbell L, King N, et al. Self-harm and suicidal content online, harmful or helpful? A systematic review of the recent evidence. JPMH 2022;21(1):57-69. [doi: 10.1108/jpmh-09-2021-0118]
34. Soldaini L, Walsh T, Cohan. Helping or Hurting? Predicting changes in users' risk of self-harm through online community interactions. 2018 Presented at: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic; 2025 October 27; New Orleans, LA p. 194-203. [doi: 10.18653/v1/w18-0621]

35. Razi A, Alsoubai A, Kim S, Ali S, Stringhini G, De Choudhury M, et al. Sliding into My DMs: detecting uncomfortable or unsafe sexual risk experiences within instagram direct messages grounded in the perspective of youth. Proc. ACM Hum.-Comput. Interact 2023;7(CSCW1):1-29. [doi: 10.1145/3579522]
36. Atmakuru A. Artificial intelligence-based suicide prevention and prediction: a systematic review (2019-2023). Inf. Fusion 2024:102673. [doi: 10.2139/ssrn.4863171]
37. Ali S. Youth, social media, and online safety: a holistic approach towards detecting and mitigating risks in online conversations. Dissertation. Boston University. 2024. URL: https://open.bu.edu/items/33f8cd88-5c22-4932-aefb-1edfcdec9916 [accessed 2025-12-22]
38. Chancellor S, Baumer EPS, De Choudhury M. Who is the "Human" in human-centered machine learning. Proc. ACM Hum.-Comput. Interact 2019;3(CSCW):1-32. [doi: 10.1145/3359249]
39. Waller G, Newbury-Birch D, Simpson D, Armstrong E, James B, Chapman L, et al. The barriers and facilitators to the reporting and recording of self-harm in young people aged 18 and under: a systematic review. BMC Public Health 2023;23(1):158 [FREE Full text] [doi: 10.1186/s12889-023-15046-7] [Medline: 36694149]
40. Kim S, Razi A, Stringhini G, Wisniewski PJ, De Choudhury M. You don't know how i feel: insider-outsider perspective gaps in cyberbullying risk detection. ICWSM 2021;15:290-302. [doi: 10.1609/icwsm.v15i1.18061]
41. Nordin N, Zainol Z, Mohd Noor MH, Chan LF. An explainable predictive model for suicide attempt risk using an ensemble learning and Shapley Additive Explanations (SHAP) approach. Asian J Psychiatr 2023;79:103316. [doi: 10.1016/j.ajp.2022.103316] [Medline: 36395702]
42. Naseem U, Khushi M, Kim J, Dunn AG. Hybrid text representation for explainable suicide risk identification on social media. IEEE Trans. Comput. Soc. Syst 2024;11(4):4663-4672. [doi: 10.1109/tcss.2022.3184984]
43. Toliya NP, Nagarathna N. Leveraging online social content for early detection of suicidal ideation: a multi-modal deep learning approach. : IEEE; 2024 Presented at: International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications; 2024 April 22-23; Bengaluru, India p. 1-7. [doi: 10.1109/icetcs61022.2024.10544279]
44. Razi A. Instagram data donation: a case study on collecting ecologically valid social media data for the purpose of adolescent online risk detection. 2022 Presented at: CHI '22: CHI Conference on Human Factors in Computing Systems; 2022 29 April- 5 May; New Orleans LA USA p. 1-9. [doi: 10.1145/3491101.3503569]
45. Cole R. Inter-rater reliability methods in qualitative case study research. Sociological Methods & Research 2023;53(4):1944-1975. [doi: 10.1177/00491241231156971]
46. Robinson J, Witt K, Lamblin M, Spittal MJ, Carter G, Verspoor K, et al. Development of a self-harm monitoring system for victoria. Int J Environ Res Public Health 2020;17(24):9385 [FREE Full text] [doi: 10.3390/ijerph17249385] [Medline: 33333970]
47. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese bert-networks. 2019 Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2025 October 27; Hong Kong, China p. 3982-3992. [doi: 10.18653/v1/d19-1410]
48. Kobayashi S. Contextual augmentation: data augmentation by words with paradigmatic relations. 2018 Presented at: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers); 2025 October 27; New Orleans, Louisiana p. 452-457. [doi: 10.18653/v1/n18-2072]
49. Ansari G, Garg M, Saxena C. Data augmentation for mental health classification on social media. 2021 Presented at: Proceedings of the 18th International Conference on Natural Language Processing (ICON); 2025 October 27; National Institute of Technology Silchar, Silchar, India p. 152-161.
50. Coppersmith G, Leary R, Crutchley P, Fine A. Natural language processing of social media as screening for suicide risk. Biomed Inform Insights 2018;10:1178222618792860 [FREE Full text] [doi: 10.1177/1178222618792860] [Medline: 30158822]
51. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: charleston adv. 2019 Presented at: Proceedings of NAACL-HLT; 2019 June 2 - June 7; Minneapolis, Minnesota p. 8-10. [doi: 10.5260/chara.21.2.8]
52. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv 2019 [FREE Full text]
53. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. Mentalbert: publicly available pretrained language models for mental healthcare. : European Language Resources Association; 2022 Presented at: Proceedings of the Thirteenth Language Resources and Evaluation Conference; 2025 October 27; Marseille, France p. 7184-7190. [doi: 10.1016/b978-0-323-90118-5.00006-0]
54. Kodati D, Tene R. Identifying suicidal emotions on social media through transformer-based deep learning. Appl Intell 2022;53(10):11885-11917. [doi: 10.1007/s10489-022-04060-8]
55. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]

56. Chung J, Gulcehre C, Cho K, Bengio Y. Gated feedback recurrent neural networks. 2015 Presented at: Proceedings of the 32nd International Conference on Machine Learning; 2015; Proceedings of Machine Learning Research p. 2067-2075 URL: https://proceedings.mlr.press/v37/chung15.html
57. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. 2016 Presented at: KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 August 13 - 17; San Francisco California USA p. 785-794. [doi: 10.1145/2939672.2939785]
58. Boot P, Zijlstra H, Geenen R. The Dutch translation of the linguistic inquiry and word count (LIWC) 2007 dictionary. DuJAL 2017;6(1):65-76. [doi: 10.1075/dujal.6.1.04boo]
59. Hutto C, Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. ICWSM 2014;8(1):216-225. [doi: 10.1609/icwsm.v8i1.14550]
60. Ramos JE. Using TF-IDF to determine word relevance in document queries. Computer Science 2003 [FREE Full text]
61. Jain SM. Hugging face. In: Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems. Cham: Springer; 2022:51-67.
62. Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S. The ‘K’ in K-fold cross validation. 2012 Presented at: ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning; 2012 April 25-27; Bruges (Belgium) p. 441-446.
63. Ying X. An overview of overfitting and its solutions. J. Phys.: Conf. Ser 2019;1168(2):022022. [doi: 10.1088/1742-6596/1168/2/022022]
64. Mehdary A, Chehri A, Jakimi A, Saadane R. Hyperparameter optimization with genetic algorithms and XGBoost: a step forward in smart grid fraud detection. Sensors (Basel) 2024;24(4):1230 [FREE Full text] [doi: 10.3390/s24041230] [Medline: 38400385]
65. Guo Y, Graber A, McBurney RN, Balasubramanian R. Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms. BMC Bioinformatics 2010;11:447 [FREE Full text] [doi: 10.1186/1471-2105-11-447] [Medline: 20815881]
66. Hu Z, Piet J, Zhao G, Jiao J, Wagner D. Toxicity detection for free. arXiv 2025 [FREE Full text]
67. Frenda S. Sarcasm and implicitness in abusive language detection: a multilingual perspective. Universitat Politècnica de València 2022. [doi: 10.4995/thesis/10251/184015]
68. St»hle L, Wold S. Analysis of variance (ANOVA). Chemometrics and Intelligent Laboratory Systems 1989;6(4):259-272. [doi: 10.1016/0169-7439(89)80095-4]
69. Franke TM, Ho T, Christie CA. The chi-square test often used and more often misinterpreted. American Journal of Evaluation 2011;33(3):448-458. [doi: 10.1177/1098214011426594]
70. Sweeney C, Ennis E, Mulvenna MD, Bond R, O'Neill S. Insights derived from text-based digital media, in relation to mental health and suicide prevention, using data analysis and machine learning: systematic review. JMIR Ment Health 2024;11:e55747 [FREE Full text] [doi: 10.2196/55747] [Medline: 38935419]
71. Kim S, Razi A, Alsoubai A, Wisniewski PJ, De Choudhury M. Assessing the impact of online harassment on youth mental health in private networked spaces. ICWSM 2024;18:826-838. [doi: 10.1609/icwsm.v18i1.31355]
72. Alsoubai A, Razi A, Agha Z, Ali S, Stringhini G, De Choudhury M, et al. Profiling the offline and online risk experiences of youth to develop targeted interventions for online safety. Proc. ACM Hum.-Comput. Interact 2024;8(CSCW1):1-37. [doi: 10.1145/3637391]
73. Benjamins C, Eimer T, Schubert F, Mohan A, Döhler S, Biedenkapp A, et al. Contextualize me--the case for context in reinforcement learning. arXiv:2202.04500 2022 [FREE Full text]
74. Rowley P, McDermott E, Hodge S. Responding to young people who disclose self-harm: a discourse analysis of an on-line counselling service. PhD thesis. United Kingdom: Lancaster University; 2019. URL: https://eprints.lancs.ac.uk/id/eprint/145622/ [accessed 2025-12-22]
75. Galasiński D, Ziółkowska J. The end of ambivalence. A narrative perspective on ambivalence in the suicidal process. Suicide Life Threat Behav 2024;54(5):888-899. [doi: 10.1111/sltb.13101] [Medline: 38847574]
76. Fincham FD, Grych JH, Osborne LN. Does marital conflict cause child maladjustment? Directions and challenges for longitudinal research. Journal of Family Psychology 1994;8(2):128-140. [doi: 10.1037//0893-3200.8.2.128]
77. Gentner D. Spatial metaphors in temporal reasoning. In: Spatial Schemas and Abstract Thought. Cambridge, Massachusetts: The MIT Press; 2001:203-222.
78. Wray A, Grace GW. The consequences of talking to strangers: evolutionary corollaries of socio-cultural influences on linguistic form. Lingua 2007;117(3):543-578. [doi: 10.1016/j.lingua.2005.05.005]
79. Alharbi R, Langer S, Hunter C, Husain N, Varese F, Taylor PJ. "My entire life has moulded the person that I am": narrations of non-suicidal self-injury and complex trauma in individuals with complex posttraumatic stress experiences. Psychol Psychother 2025;98(3):683-700. [doi: 10.1111/papt.12583] [Medline: 40022593]
80. De Choudhury M, Kiciman E. The language of social support in social media and its effect on suicidal ideation risk. ICWSM 2017;11(1):32-41. [doi: 10.1609/icwsm.v11i1.14891]
81. Yıldırım N. The bullying game: sexism based toxic language analysis on online games chat logs by text mining. Journal of International Women's Studies 2022;24(3) [FREE Full text]

82. Cameron EL, Fox JD, Anderson MS, Cameron CA. Resilient youths use humor to enhance socioemotional functioning during a day in the life. Journal of Adolescent Research 2010;25(5):716-742. [doi: 10.1177/0743558410366595]
83. Kerr DCR, Preuss LJ, King CA. Suicidal adolescents' social support from family and peers: gender-specific associations with psychopathology. J Abnorm Child Psychol 2006;34(1):103-114. [doi: 10.1007/s10802-005-9005-8] [Medline: 16502141]
84. Tanni TI, Akter M, Anderson J, Amon MJ, Wisniewski PJ. Examining the unique online risk experiences and mental health outcomes of lgbtq+ versus heterosexual youth. 2024 Presented at: CHI '24: CHI Conference on Human Factors in Computing Systems; 2024 May 11 - 16; Honolulu HI USA p. 1-21. [doi: 10.1145/3613904.3642509]
85. Arık E, İnce M, Koçak MC, Bilişli Y, Karataş EO, Akgün H, et al. Communication dynamics and media interactions of young adults who have attempted suicide: a qualitative thematic analysis. Front Psychol 2024;15:1460348 [FREE Full text] [doi: 10.3389/fpsyg.2024.1460348] [Medline: 39554706]
86. Klineberg E, Kelly MJ, Stansfeld SA, Bhui KS. How do adolescents talk about self-harm: a qualitative study of disclosure in an ethnically diverse urban population in England. BMC Public Health 2013;13:572 [FREE Full text] [doi: 10.1186/1471-2458-13-572] [Medline: 23758739]
87. Mohammar K, Windahl J. Potentially harmful, probably harmless defining and reducing false positives in digital risk assessment. Dissertation.: Uppsala University; 2022. URL: https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-497930 [accessed 2025-07-27]
88. Bentley KH, Zuromski KL, Fortgang RG, Madsen EM, Kessler D, Lee H, et al. Implementing machine learning models for suicide risk prediction in clinical practice: focus group study with hospital providers. JMIR Form Res 2022;6(3):e30946 [FREE Full text] [doi: 10.2196/30946] [Medline: 35275075]
89. Crisis text line. Home Page. 2025. URL: https://www.crisistextline.org/ [accessed 2025-07-27]
90. Dynamic, experiential learning proven to change student lives. Kognito. 2025. URL: https://kognito.com/ [accessed 2025-07-27]
91. Alsoubai A, Park JK, Stringhini G, Ma M, De Choudhury M, Wisniewski PJ. Timeliness matters: leveraging reinforcement learning on social media data to prioritize high-risk conversations for promoting youth online safety. 2025 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; 23-26 June, 2025; Copenhagen, Denmark p. 37-51 URL: https://ojs.aaai.org/index.php/ICWSM/article/view/35802 [doi: 10.1609/icwsm.v19i1.35802]
92. Morken IS, Dahlgren A, Lunde I, Toven S. The effects of interventions preventing self-harm and suicide in children and adolescents: an overview of systematic reviews. F1000Res 2019;8:890. [doi: 10.12688/f1000research.19506.1]
93. Marks M. Artificial intelligence based suicide prediction. 21 Yale Journal of Law and Technology 98 2019 [FREE Full text]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
BERT: Bidirectional Encoder Representations from Transformers
DistilBERT: Distilled Bidirectional Encoder Representations from Transformers
IGDD: Instagram Data Donation
LIWC: linguistic injury word count
ML: machine learning
NLP: natural language processing
RQ: research question
SBERT: Sentence-Bidirectional Encoder Representations from Transformers
SH-S: self-harm and suicide
SVM: support vector machine
XGBoost: extreme gradient boosting

Edited by J Torous; submitted 15.Apr.2025; peer-reviewed by S Saripalli, T Kayode, SB Guo, M Elbattah; comments to author 03.Jun.2025; revised version received 29.Jul.2025; accepted 01.Aug.2025; published 23.Jan.2026.

Please cite as:
Qadir S, Alsoubai A, Park JK, Ali NS, Choudhury MD, Wisniewski P
Triaging Casual From Critical—Leveraging Machine Learning to Detect Self-Harm and Suicide Risks for Youth on Social Media: Algorithm Development and Validation Study
JMIR Ment Health 2026;13:e76051
URL: https://mental.jmir.org/2026/1/e76051
doi:10.2196/76051
PMID:

Original Paper

“It Felt Good to Be Able to Say That Out Loud”—Therapeutic Alliance and Processes in AVATAR Therapy for People Who Hear Distressing Voices: Peer-Led Qualitative Study

Emily Rutter-Eley[1,2], MSci, DClinPsy; Thomas Craig[2,3], MBBS, FRCP, PhD; Philippa Garety[1,2], MPhil, PhD, MA; Mar Rus-Calafell[1,4], ClinPsyD, PhD; Hannah Ball[5,6], DClinPsy; Moya Clancy[7,8], MSc, DClinPsy; Jeffrey McDonnell[9,10], MSc, DClinPsy; Andrew Gumley[7,8], PhD, MAppSci, CPsychol, AFBPsS; Gillian Haddock[5,6], MClinPsychol, PhD; Sandra Bucci[5,6], ClinPsyD; Miriam Fornells-Ambrojo[9,10], DClinPsy, PhD; Nerys Baldwin[11]; Jed Harling[11]; Alie Phiri[11], MSc; Charlie MacKenzie-Nash[11]; Nicholas Hamilton[11]; Amy Grant[11]; Clementine Edwards[1,2*], MSc, DClinPsy, PhD; Thomas Ward[1,2*], MSci, DClinPsy, PhD

[1]Department of Psychology, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom
[2]South London & Maudsley NHS Foundation Trust, London, United Kingdom
[3]Department of Health Service and Population Research, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom
[4]Mental Health Research and Treatment Center, Faculty of Psychology, Ruhr University Bochum, Bochum, Germany
[5]Division of Psychology and Mental Health, School of Health Sciences, University of Manchester and the Manchester Academic Health Sciences Centre, Manchester, United Kingdom
[6]Greater Manchester Mental Health NHS Foundation Trust and the Manchester Academic Health Sciences Centre, Manchester, United Kingdom
[7]School of Health and Wellbeing, University of Glasgow, Glasgow, United Kingdom
[8]NHS Greater Glasgow and Clyde, Glasgow, United Kingdom
[9]Research Department of Clinical, Educational and Health Psychology, University College London, London, United Kingdom
[10]North East London NHS Foundation Trust, London, United Kingdom
[11]The AVATAR2 PPI Group, London, United Kingdom
[*]these authors contributed equally

Corresponding Author:
Thomas Ward, MSci, DClinPsy, PhD
Department of Psychology
Institute of Psychiatry, Psychology & Neuroscience
King's College London
16 De Crespigny Park
London, SE5 8AB
United Kingdom
Phone: 44 020 7848 0002
Email: thomas.ward@kcl.ac.uk

Abstract

Background: AVATAR therapy is a novel psychological therapy that aims to reduce distress associated with hearing voices. The approach involves a series of therapist-facilitated dialogues between a voice-hearer and a digital embodiment of their main distressing voice (the avatar), which aim to increase coping and self-empowerment.

Objective: This study explored therapeutic processes that are distinctive to AVATAR therapy, including direct early work with voice content and the role of the therapist in dialogue enactment.

Methods: People with lived experience relating to psychosis (peer researchers) contributed to each stage of the study. Peer researchers led semistructured interviews, which were conducted with 19 participants who received AVATAR therapy as part of the AVATAR2 trial, including 3 participants who dropped out of therapy. Data were analyzed using interpretative phenomenological analysis (n=5) and template analysis (n=14).

Results: Participants described the initial challenges of experiential work with distressing voice content; however, most reported a meaningful increase in power and control over the course of dialogues and improvements with voices in daily life. A strong therapeutic alliance was experienced by all participants, including those who chose to discontinue therapy, often mitigating the discomfort associated with initial challenges by enhancing their sense of safety. Several important themes relating to individual engagement were highlighted, such as the emotional intensity of the experience and the importance of participants' determination and open-minded attitudes despite initial doubts. Those who decided not to continue with therapy described challenges with the realism of working dialogically with a digital representation of their distressing voice.

Conclusions: This study has provided a deeper understanding of the experience of engaging in AVATAR therapy, in particular the challenges and opportunities of direct work with voice content. The importance of therapeutic alliance and establishing a sense of voice presence has been emphasized. Implications for the planned optimization and wider implementation of AVATAR therapy in routine care settings are discussed.

Trial Registration: ISRCTN Registry ISRCTN55682735; https://www.isrctn.com/ISRCTN55682735

(JMIR Ment Health 2026;13:e77566) doi:10.2196/77566

KEYWORDS

auditory hallucinations; AVATAR therapy; peer research; psychosis; qualitative study

Introduction

Voice-hearing is a diverse experience that can occur in both clinical and nonclinical populations. While some voices may be experienced as neutral or positive, others are distressing, hostile, and disruptive to daily life [1,2]. Recent developments in psychological interventions for distressing voices have foregrounded the relationship between the voice-hearer and voice as a key treatment target [3-5]. AVATAR therapy is one such approach, which aims to reduce the distress associated with hearing negative voices [6]. In AVATAR therapy, voice-hearers are supported to create the voice and image of a digital avatar to represent their main distressing voice. There is flexibility to customize facial and vocal features, including nonhuman forms (eg, devil), and the avatar is voiced in real time by the therapist using voice-transformation software. The person interacts with the avatar in a series of dialogues, supported by the therapist and interspersed with preparatory and reflective discussions (including role play), with the aim of taking back power and control from the voice [7,8]. Further details on therapeutic processes and targets are described in the main trial [9] and Ward et al [8]. Early work [7] identified 2 therapy phases. The focus of phase 1 is on exposure to the avatar voice enactment (including verbatim voice content) and encouraging the voice-hearer to respond assertively. In phase 2, the avatar concedes power to the voice-hearer, and the focus shifts to other therapy targets, including self-esteem.

Building on Leff et al [6] work, this approach was tested in a fully powered randomized controlled trial (RCT; AVATAR1), which found that AVATAR therapy was more effective than supportive counseling in reducing the frequency, distress, and omnipotence of voices at 12 weeks post therapy with a large between-group effect size of 0.8 for overall voice severity [7]. Following this, a multicenter RCT (AVATAR2 [9]) has recently tested a 6-session version of AVATAR therapy (AVATAR-Brief [AV-BRF], comprising the key elements of exposure, assertiveness, and self-esteem) alongside an extended (12-session) form of AVATAR therapy (AV-EXT). The AV-EXT protocol was designed to foreground the understanding of voices within the person's broader life history [10] and include a wider range of therapeutic targets, including trauma. The AVATAR2 trial found that voice distress (primary outcome) and voice severity significantly improved in both AV-BRF and AV-EXT at the end of therapy, compared to treatment as usual alone, with improvements maintained at 28 weeks but no longer statistically significant. There was a significant and sustained reduction in voice frequency for AV-EXT, but not AV-BRF. While both therapy groups showed significant improvements across a range of other secondary outcomes, AV-EXT showed a wider range of positive effects in areas including increased empowerment, voice understanding, and well-being, and these tended to be stronger and longer lasting [9].

While AVATAR therapy shares common therapeutic aims with other cognitive and relational therapies, there are several distinctive aspects, notably direct exposure to feared stimuli and real-time dialogues with a digital embodiment of the voice enacted by the therapist [8,11]. This includes the experience of "voice presence," defined as the degree to which the dialogue with the avatar is experienced "as if" talking to the voice [12,13]. These unique features raise important questions around the blending of digital technology and relational therapy, particularly regarding the experience of therapeutic alliance (TA) within this distinct digital context, which differs from app-based interventions discussed in existing literature [14-17]. The participant experience of AVATAR therapy was examined within a previous qualitative study conducted as part of the AVATAR1 trial [11]. The use of technology was generally well accepted by participants, and the collaborative process of designing the avatar and enacting the relationship with the voice was reported to be helpful in facilitating therapeutic dialogue. Participants described feeling supported by their therapist and were able to identify specific learned strategies for managing voices, such as standing up to the voice and choosing to disengage. Although Rus-Calafell et al [11] provided important support for the acceptability of AVATAR therapy, limitations included the recruitment of only 1 (5%) therapy dropout and the inclusion of patient and public involvement (PPI) into some (eg, developing topic guide, conducting interviews) but not all research processes. It was recommended that future qualitative work should aim to build understanding of the specific challenges of early direct work with derogatory verbatim voice content given that this represents a central aspect of the AVATAR therapy approach and a potential point of difference from other psychological approaches.

This study differed in that it used interpretative phenomenological analysis (IPA), a qualitative approach that aims to understand the emotions and personal meanings associated with experiences in the context of an individual's lifeworld [18,19]. IPA was selected to provide in-depth, idiographic insights into participants' subjective experiences [20]. Template analysis is a form of thematic analysis that involves developing and refining a hierarchical coding template, often beginning with a priori codes, which is iteratively applied and modified across the dataset to organize themes in a meaningful way [21]. As Frost [22] highlights, interpreting data pluralistically using multiple methods of analysis can facilitate a more unbiased and holistic perspective of participants' experiences. Template analysis can be incorporated alongside IPA to extend the scope of the analysis beyond an individual case level, enabling identification of convergences and divergences across participants, thus providing a more comprehensive understanding [23]. This dual approach of IPA and template analysis was used by Bond et al [24] to capture both the nuances of individual experiences and the broader shared themes, ensuring both depth and breadth [23]. Additionally, this analytic approach recognizes the potential influence of the researcher on interpretation [25], highlighting the value of a peer research approach to improve both the rigor and relevance of qualitative analysis [26]. Therefore, involving people with relevant lived experience, as demonstrated by Bond et al [24], can enrich the depth, sensitivity, and ecological validity of the analysis [27].

In this study, we aim to move beyond a validation of the acceptability of the approach to provide a rich investigation of therapeutic features that are distinctive to AVATAR therapy. Using peer methods across all stages of the study and improving representation of those who dropped out of therapy, we aim to address the following research questions:

- How did participants experience working directly with verbatim voice content?
- How was TA experienced by participants in the context of AVATAR therapy?
- What were the experiences of participants who decided not to continue with therapy?

Methods

Study Design and Setting

This qualitative study investigated participant experiences of AVATAR therapy based on semistructured interviews. Adopting a peer research approach, this study was nested within the AVATAR2 trial (ISRCTN55682735, registered on January 22, 2020), a multicenter RCT conducted across 4 trial centers in England and Scotland (Institute of Psychiatry, Psychology, and Neuroscience, King's College London; University College London [UCL]; The University of Manchester; and University of Glasgow). The additional research ethics approval for this study was granted by the London-Camberwell St Giles Research Ethics Committee in December 2022 (reference 20/LO/0657).

In line with open science processes, a preregistration form (based on Haven et al [28] and Staniszewska et al [29]) detailing the study protocol was published on the Open Science Framework to ensure transparency and reliability [30].

Patient and Public Involvement

A total of 16 PPI representatives (termed peer researchers in this study) were involved across all stages of this study, with representation from all 4 trial sites. Most individuals had lived experience of hearing distressing voices, with some having received AVATAR therapy as part of the AVATAR1 or AVATAR2 trial, as well as a small number of people who had cared for a loved one who had experienced this. This included (1) coproduction of participant-facing documentation, (2) collaborative development of the topic guide, (3) codelivery of interviews, and (4) contributions to data analysis. This peer research approach aligns with the National Institute for Health Research INVOLVE guidelines [31], which highlight 6 key values—respect, support, transparency, responsiveness, fairness of opportunity, and accountability. For example, all peer researchers were paired with a site research worker who provided flexible and personalized support in order to develop a collaborative working relationship [32]. Peer researchers were remunerated according to INVOLVE guidelines [31]. More details are provided in Multimedia Appendix 1, and PPI across the trial is further discussed in Owrid et al [32].

Participants

We recruited participants who had received AVATAR therapy as part of the AVATAR2 trial [33]. This included those randomized to the AV-BRF and AV-EXT arms. All participants met eligibility criteria for the trial. Inclusion criteria were (1) aged ≥18 years, (2) under the care of a mental health team, (3) hearing a distressing voice in English for at least 6 months, (4) adequate English to take part, and (5) schizophrenia spectrum disorder or an affective disorder with psychotic symptoms. Exclusion criteria were (1) lacking capacity to consent, (2) currently undertaking individual psychological therapy for voices, and (3) currently experiencing an acute mental health crisis. In addition, specific inclusion criteria for this study were as follows: (1) received AVATAR therapy in the AVATAR2 trial, (2) able and willing to provide informed consent to take part in the interview, and (3) willing to have the interview audio-recorded.

Sampling and Recruitment

As IPA principles require homogeneity around a phenomenon of interest [19,34], this was implemented in terms of adults experiencing distressing voices who have received AVATAR therapy. A predetermined recruitment strategy was consistently followed by trial coordinators at each site. This involved initially using consecutive sampling methods to invite participants for interview who had recently completed the final phase of the AVATAR2 trial. To achieve representativeness across sample characteristics and therapy delivery, purposive sampling methods were then applied to ensure that (1) the sample represented all 4 trial sites, (2) the gender and ethnicity of the sample were representative of participants across the trial, (3) the sample represented both AV-BRF and AV-EXT therapy arms, (4) the interviews covered experiences of working with a range of therapists, and (5) those recruited included a proportionate number of those who did not complete the intervention.

The aim was to recruit 20 participants across the 4 trial sites, including 4 participants who dropped out of therapy.

Interview Guide

Semistructured interviews were guided by a topic guide (Multimedia Appendix 2) that was developed iteratively in collaboration with peer researchers, trial coordinators, therapy leads, and site research workers. Key areas of questioning were based on the research questions and initially drafted by the study team in line with IPA principles, such as focusing on the individual's emotional experiences and associated meanings. The guide was then revised across 4 consultations, involving a total of 5 peer researchers. For example, questions were reworded to ask about "the voices" instead of "your voices" to ensure destigmatizing and nonjudgmental language. Interview questions explored participants' (1) experience of dialoguing with a representation of their distressing voice, (2) TA formed with the therapist, and, where applicable, (3) reasons for not completing therapy.

Interview Process

Interviews were conducted either in person, online via Microsoft Teams, or via telephone, depending on participant preference. All interviews were conducted by a site research worker and peer researcher. In line with Harding et al [35], peer researchers took the lead in interviewing all participants while the accompanying site research worker provided support and input as required. Peer researchers disclosed their relevant lived experience when introducing themselves, but it was at their discretion how much detail they shared during the interview. A total of 14 peer researchers were involved in leading interviews, all of whom attended training (Multimedia Appendix 3), role-play practice, and supervision. The lead author, ERE, was available to peer researchers, site research workers, and trial coordinators for support with all aspects of protocol delivery to ensure a standardized approach.

Analysis

Data were analyzed using a dual approach of IPA and template analysis, as described by Bond et al [24,36]. Nineteen is a relatively large sample for IPA's rigorous and idiographic approach. Therefore, a small core sample of 5 (26.3%) transcripts was included in IPA processes, with analysis then extended across the remaining transcripts using template analysis. In line with IPA principles, the selection of transcripts was based on the richness of data and ensured diversity in trial sites, gender, and ethnicity. To ensure homogeneity, as required for IPA [19,34], all selected transcripts were from therapy completers. This selection process was led by ERE with guidance from experts in qualitative research methods and input from peer researchers who conducted interviews.

IPA followed analytic processes described by Smith et al [18] and Smith and Nizza [37]. The 4 peer researchers who conducted the 5 selected interviews first listened to their interview recordings to reflect on their emotional responses and perspectives of participants' experiences. ERE then met individually with each peer researcher in 1-hour consultations to discuss interpretations. Using a case-by-case approach, line-by-line annotations were made for each transcript, incorporating the peer researcher's perspectives and noting descriptive, linguistic, and conceptual comments. Identified meanings were therefore informed by peer researchers' interpretations and then formulated into experiential statements, which were subsequently clustered and interpreted for each case. Convergences and divergences of themes were then considered across cases, generating group experiential themes and subthemes. Excerpts of the analytic process are presented in Multimedia Appendix 4.

Before applying template analysis, the resulting coding structure (ie, group experiential themes and subthemes) alongside illustrative quotes was reviewed by 6 peer researchers across 4 consultations to develop the provisional template (Multimedia Appendix 5). This input extended initial interpretations by drawing on lived experience, for example, emphasizing the complexity of the avatar-voice association, highlighting the profound nature of participants' emotional experiences, and helping to identify language that felt authentic. This template was then used to guide template analysis of the remaining transcripts, following the processes outlined by King et al [21]. Transcripts were introduced in subgroups, focusing first on therapy completers and, second, on those who dropped out of therapy. This clearly identified where these 2 subgroups aligned and where new experiences arose. Where data diverged, new codes were assigned, and the template was changed to accommodate these perspectives. Once all transcripts had been coded and the template had been developed iteratively, all transcripts were reviewed again to ensure reliability.

Furthermore, 4 peer researchers reviewed the finalized template and contributed to refining codes, wording themes, and interpreting results. For example, the subtheme initially named "Difficulties generalizing learning" was reworded to "Difficulties translating changes to voices" to ensure accessible, nonclinical language. Additionally, the subthemes "Initial doubts and anxieties" and "Positive, open-minded attitudes" were merged into "Open-minded attitudes despite initial doubts" to reflect peer researchers' interpretation that open-mindedness acted as an overriding force against early anxieties.

IPA and template analyses were conducted by ERE with input from peer researchers across the analytic process, ensuring that lived experience actively contributed to meaning-making. The outcome of this dual approach was a structure of key themes and subthemes (ie, the finalized template) representing all 19 interviews, including the 5 transcripts analyzed using IPA.

Reflexivity

All members of the AVATAR2 team reflected on the perspectives they brought to the study design, conduct, and analysis through whole team meetings, peer group supervision with site research workers and peer researchers, and the lead author's reflective research journal. The team approach and wider context of the qualitative research program were published on the Open Science Framework [30]. The lead author (ERE), a White British female trainee clinical psychologist, had no involvement in the AVATAR1 trial nor in the wider AVATAR2 trial, as her role was limited to this qualitative study, and had no prior contact with participants. While her clinical experience of working with people with psychosis, including delivering cognitive behavioral therapy, may have fostered positive attitudes toward psychological therapy, she had no clinical experience specifically with AVATAR therapy. Completing the Jacobson and Mustafa [38] positionality map supported reflection on social identity and potential sources of bias.

Ethical Considerations

This study was approved by the London-Camberwell St Giles Research Ethics Committee in December 2022 (reference 20/LO/0657) and was conducted in accordance with the principles of the Declaration of Helsinki.

The participants in this qualitative study signed informed consent before participating in the interview that included consent for publication. All potentially identifiable information has been removed from published material included in this study. All interviews were audio-recorded, transcribed, and anonymized. Participants were paid £20 (US $26.65) for taking part in the interview.

Results

Participants

A total of 19 participants took part in qualitative interviews, which were conducted between July and September 2023 and took place between 11.4 and 96.4 (mean 53, SD 29.1) weeks after their final AVATAR therapy session. Overall, 6 (31.6%) participants were recruited from King's College London, 4 (21.1%) from UCL, 4 (21.1%) from Manchester, and 5 (26.3%) from Glasgow. Demographic and treatment information of participants are presented in Table 1 alongside that for the whole study sample of those who received AVATAR therapy (AV-BRF and AV-EXT). Participants in the interview sample were broadly representative of the wider sample; however, differences were not statistically tested.

Table 1. Demographic and treatment information.

Variable	Whole sample (n=230)	Interview sample (n=19)
Age (years), mean (SD; range)	40.07 (13.49; 18-70)	38.63 (14.06; 19-66)
Gender, n (%)		
Men	143 (62.2)	13 (68.4)
Women	85 (37)	6 (31.6)
Other	2 (0.9)	0 (0)
Ethnicity, n (%)		
White	136 (59.1)	11 (57.9)
Black Caribbean	13 (5.7)	0 (0)
Black African	21 (9.1)	2 (10.5)
Black Other	7 (3)	0 (0)
Indian	5 (2.2)	1 (5.3)
Pakistani	8 (3.5)	0 (0)
Chinese	2 (0.9)	1 (5.3)
Other	38 (16.5)	4 (21.1)
Treatment arm, n (%)		
Brief	116 (50.4)	10 (52.6)
Extended	114 (49.6)	9 (47.4)
Dropout, n (%)		
Yes	69 (30)	3 (15.8)
No	161 (70)	16 (84.2)

A total of 9.9% (16/161) of therapy completers and 4.3% (3/69) of those who discontinued therapy are represented in the sample. Of those who dropped out of therapy, 2 participants decided not to continue after 1 session prior to any avatar dialogue, and 1 participant dropped out after 8 sessions (including 5 avatar dialogues). The 3 participants who dropped out, Grace, James, and Alexander (pseudonyms), were all allocated to AV-EXT.

Key Findings

Integrating IPA and template analysis processes, 4 overarching themes were identified, from which a total of 14 subthemes emerged. Details of themes alongside the number of participants are represented in Table 2 and subsequently presented in more depth. This is summarized in Multimedia Appendix 6.

Illustrative quotes are presented within the text and in Multimedia Appendix 7. Pseudonyms are used throughout.

Table 2. Themes and subthemes identified, with participant counts.

Themes and subthemes	Sample, n (%)	Completers, n (%)	Noncompleters, n (%)
Shift in relationship with avatar and, consequently, voices			
Initial challenges adjusting to avatar	18 (94.7)	15 (93.8)	3 (100)
Collaborative efforts facilitated meaningful connection to avatar	14 (73.7)	13 (81.3)	1 (33.3)
With therapist support, participants felt empowered to stand up to avatar	16 (84.2)	15 (93.8)	1 (33.3)
Positive shift with voices	15 (78.9)	13 (81.3)	2 (66.7)
Crucial role of person-centered therapist			
Felt safe, supported, and understood	19 (100)	16 (100)	3 (100)
Person-centered flexibility	15 (78.9)	13 (81.3)	2 (66.7)
Significant impact of therapeutic alliance	16 (84.2)	14 (87.5)	2 (66.7)
Individual approach and experience			
Open-minded attitudes despite initial doubts	18 (94.7)	15 (93.8)	3 (100)
Determination facilitated engagement and outcomes	17 (89.5)	16 (100)	1 (33.3)
Profound emotional experience	17 (89.5)	14 (87.5)	3 (100)
Offered novel approach to tackle voices	12 (63.2)	11 (68.8)	1 (33.3)
Barriers to engagement and outcomes			
Emotional challenges with avatar	10 (52.6)	7 (43.8)	3 (100)
Not the right approach for the individual at that time	8 (42.1)	5 (31.3)	3 (100)
Difficulties translating changes to voices	6 (31.6)	6 (37.5)	0 (0)

Theme 1: Shift in Relationship With Avatar and, Consequently, Voices

This theme explores how participants' experiences of working dialogically with verbatim voice content evolved across therapy. The subthemes "Initial challenges adjusting to avatar" and "Collaborative efforts facilitated meaningful connection to avatar" focus on initial adjustment processes. Subsequently, the subtheme "With therapist support, participants felt empowered to stand up to avatar" explores communication in dialogues, and the final subtheme, "Positive shift with voices," highlights improvements in coping.

Initial Challenges Adjusting to Avatar

Being exposed to derogatory voice content through the avatar could be challenging at first, with initial reactions ranging from fear to emotional disconnection. For 5 participants, the avatar immediately provided an accurate representation and felt real, evoking a similar emotional response to hearing voices. This could lead to heightened anxiety.

> *That was probably one of the most difficult days of my life, that very first session [...] I was a just a complete utter mess that day.* [Charlotte, AV-BRF, completed therapy]

Conversely, 4 participants experienced emotional disconnection due to discrepancies between the avatar and the voices. Accurately matching these elements was a common challenge, and many felt it was initially difficult to fully capture both the content and intensity of such a dynamic and individualized experience.

> *It was off-putting [...] it was harder for me to really engage with it.* [Stephen, AV-BRF, completed therapy]

This was a spectrum of experience, and most reported only short-term discomfort within early sessions. In particular, 6 participants described that it was initially strange to experience the familiarities of verbatim voice content in a new context.

> *Rather than just, you know, in my mind, it was actually there in front of me.* [Grace, AV-EXT, dropped out of therapy]

This tangible representation appeared particularly confronting for participants who had used avoidance to cope.

> *I was pretty nervous to make things that I didn't want to be real seem more real.* [Gabriel, AV-EXT, completed therapy]

Fears arose that the avatar might merge with other distressing experiences. In fact, this did occur for Gabriel, although he highlighted that this was ultimately useful for his engagement and overcoming difficulties, as it connected his experience of the avatar to his experience of voices.

> *The face and the voice of the avatar then becoming sort of combined with the other things that I was seeing and hearing at the time.* [Gabriel, AV-EXT, completed therapy]

Therefore, initial exposure to the avatar was a complex and dynamic experience. For instance, Stephen initially struggled with the software, particularly as he felt customization options for creating a Black female face and voice were limited, so felt emotionally detached from the avatar as it did not accurately represent his voice experience. However, following the first session, he noticed increased voices in daily life. While temporarily challenging, this helped him make a stronger link between the avatar and his voice experience and subsequently enhanced his emotional engagement with the avatar.

> *In the time of doing it, I didn't really feel anything from it, it felt very gimmicky [...] as soon as I'd finished the session then the voices kind of hit me and they were quite intense.* [Stephen, AV-BRF, completed therapy]

For 5 participants, these challenges with heightened anxiety and emotional disconnection impacted early dialogues, leading to feeling unable to respond to the avatar or subconsciously reverting to ignoring voice content.

> *Feelings of almost of being choked out, I couldn't really think straight.* [Ishan, AV-EXT, completed therapy]

> *Went into default mode of just ignoring it.* [Stephen, AV-BRF, completed therapy]

Collaborative Efforts Facilitated Meaningful Connection to Avatar

Adjusting and collaboratively working through initial challenges was crucial to relate to the avatar as if they were relating to the voice. Nine participants reported that increased exposure over time improved familiarity and comfort levels.

> *As I did one or two sessions, I just got used to it and [...] I knew what was going to happen.* [Zahid, AV-BRF, completed therapy]

Participants learned to accept any inaccuracies and instead focus on the experience of interacting with the avatar. Three participants described making efforts to overlook discrepancies and using their imagination to "fill in a lot of blanks" in order to relate the avatar to the voice.

> *I realised it really wasn't about being dead accurate to what the voices was sounding like but more about the interaction with it.* [Ishan, AV-EXT, completed therapy]

Therapists played a crucial role in adjusting software based on participant feedback and enacting the avatar to represent the voice both authentically and respectfully.

> *My therapist was so willing and helpful with that and presented that in a way that was respectful but also true to my experience, made it very visceral and real for me.* [Stephen, AV-BRF, completed therapy]

As verbatim voice content was often derogatory, this involved open and ongoing conversations to ensure informed consent. Individualized discussions around therapy aims and processes helped participants to understand and engage meaningfully. For Joshua, creating a story for the avatar enhanced his engagement. As the avatar represented the voice of God, his therapist supported him to frame the voice's aggression as stemming from a lack of understanding about human experience. Therefore, dialogues were used as a space for him to explain to the avatar the complexities of being human, with the aim of helping the avatar to understand him better and reduce its anger.

> *We started to kind of like develop [...] a narrative to the avatar and like understand the avatar in a way.* [Joshua, AV-BRF, completed therapy]

Collaborative efforts improved most participants' connection to the avatar, facilitating engagement and meaningful dialogues.

> *Once you get that, it's so real and it has such a profound effect that it doesn't necessarily matter whether it matches the voice.* [Stephen, AV-BRF, completed therapy]

However, challenges did persist for the 3 therapy completers who did not perceive notable benefits from therapy (Zahid, Georgia, and Mai Su), and they continued to struggle to connect the avatar to the voice.

> *Getting the two together was [...] the difficult bit.* [Zahid, AV-BRF, completed therapy]

With Therapist Support, Participants Felt Empowered to Stand Up to Avatar

Most participants experienced a gradual but distinct shift in power dynamics across therapy. Initially, many felt overpowered by the avatar, mirroring their day-to-day struggles with voices, but 14 participants described gaining confidence over time. This was frequently conceptualized as a "battle" in which participants fought back and progressively gained control.

> *It actually felt quite good to stand up to it [...] I've just let him use me as a big punch bag, but with AVATAR therapy, I felt as if I was gaining more control.* [Charlotte, AV-BRF, completed therapy]

For 4 participants, focusing on the avatar's face helped them feel more assertive, as they could direct their communication to something specific and externalized. The process of responding to the avatar was highly individualized, depending on the context and relationship with the voice. For some, it involved learning to disengage, while others focused on more compassionate communication.

> *It really wasn't about winning; it was sort of about leaving the conversation with the avatar at an agreeable sort of level.* [Ishan, AV-EXT, completed therapy]

Therapist support before, during, and after dialogues was identified as crucial in facilitating positive change. This included role-play, reassurance, and reflective discussions. The therapist's unique position as both a participant and observer within dialogues was distinctly valued in facilitating new insights.

> *I think the confidence grew basically just by practising and [...] talking after the actual avatar session with the therapist.* [Ishan, AV-EXT, completed therapy]

Positive Shift With Voices

As participants gained confidence in confronting the avatar, many observed mirroring of improvements with voices. The metaphor of a battlefield arose again as participants compared being "in the trenches" prior to therapy to now equipped to fight back, such as with "a shield to [...] block out the voices."

> *I'm able to answer them back, in a much more direct way that I used to.* [William, AV-EXT, completed therapy]

A total of 13 participants reported reductions in voice frequency and severity due to improved confidence and abilities to cope with voices. Consequently, 9 described improved quality of life, such as increased social interaction made possible by the reduced impact of voices. In addition, 8 participants reported an improved understanding of voices supported by discussions with the therapist, which helped them to accept and reframe their experiences and develop effective coping strategies. Although voices persisted for all participants, their impact notably reduced for most.

> *They're not as bad now, they kind of sit back and leave me alone.* [Caleb, AV-EXT, completed therapy]

Skills and insights gained could have a lasting impact, and 6 participants highlighted the importance of continued practice to maintain positive changes. Experiences in AVATAR therapy often had a ripple effect across participants' lives, fostering inner strengths and rebuilding self-esteem.

> *I felt empowered that like if I could get through that, then there's not really much that I can't get through.* [Stephen, AV-BRF, completed therapy]

Theme 2: Crucial Role of Person-Centered Therapist

The role of the therapist in AVATAR therapy was emphasized. The subthemes "Felt safe, supported, and understood" and "Person-centered flexibility" highlight valued aspects of the therapist's approach and qualities, and the subtheme "Significant impact of TA" explores the effect of the therapeutic relationship on engagement and outcomes.

Felt Safe, Supported, and Understood

Initial discomfort, particularly when sharing verbatim voice content, was prevalent due to fears of stigma, misunderstanding, and judgment, often compounded by negative past experiences.

> *I find it really difficult to talk about him and that's why I try and keep him in, because I don't want other people judging.* [Charlotte, AV-BRF, completed therapy]

Despite these concerns, all participants described feeling secure and supported early in therapy, which facilitated opening up. Professional boundaries, confidentiality, and therapist credentials contributed to this sense of safety.

> *I felt like it was a safe environment, it was quite enclosed and isolated from other people.* [Ishan, AV-EXT, completed therapy]

A total of 18 participants highlighted the importance of core therapeutic skills, such as empathy, transparency, and clear communication. Participants felt heard and accepted without judgment, fostering a trusting therapeutic relationship and providing a valued contrast to past experiences.

> *I could feel the empathy and she wasn't sympathetic, but her empathy was there and [...] I was able to confide in her.* [Paula, AV-BRF, completed therapy]

In addition, 13 participants also described feeling deeply understood, enhanced by therapists' specialist knowledge about voice-hearing. In fact, it was identified how therapists developed a distinct depth of understanding through direct exposure to voice content via the avatar. Gaining an ally in this way could feel validating.

> *I feel like there's only so much somebody can really understand until they're literally being face-to-face [...] to role play that with me, I feel that there's a certain level of understanding that you get from that that you wouldn't get from other interventions.* [Stephen, AV-BRF, completed therapy]

Person-Centered Flexibility

A total of 14 participants highlighted that flexibility and collaboration empowered them to engage, as they felt involved as an equal partner. Unpressured, person-centered pacing and choice were particularly valued, especially surrounding avatar dialogues.

> *Always make sure I had a say, if I was able to continue, if I wanted to stop, if I was able to do it, whatever [...] nothing was imposed on me.* [Paula, AV-BRF, completed therapy]

Central to this experience, therapists remained attuned to participants' emotional and psychological states, so they were flexible in responding to individual needs.

> *Could tell when I was not having a good day, which was important to me [...] they recognised that and were able to work around it, which really helped me.* [Matthew, AV-EXT, completed therapy]

Participants also valued therapist flexibility in offering space to discuss issues outside of avatar dialogues, which helped to explore alternative interpretations of their experiences and develop new insights.

> *He was really, really helpful and flexible and just let me speak about whatever I wanted to speak about.* [Gabriel, AV-EXT, completed therapy]

Collaborative use of other relevant materials, such as psychoeducation, was highlighted by 4 participants for its value in reducing stigma and enhancing understanding.

> *We went through a document on different types of intrusive thoughts people have and I related to quite a lot of them [...] so it does make me feel a bit more normal.* [Joshua, AV-BRF, completed therapy]

Additionally, 8 participants valued therapists being flexible in rescheduling appointments, arranging transport, offering remote options, and making check-in calls between sessions.

> *There was a flexibility [...] that allowed me to be able to complete.* [Stephen, AV-BRF, completed therapy]

It felt important for therapists to embrace the participant's individuality, understand their identity, and acknowledge unique needs, strengths, and values. This person-centered approach enhanced TA, engagement, and outcomes.

> *I was allowed to be myself even while trying to get help.* [Paula, AV-BRF, completed therapy]

> *People are so simple but yet so incredibly complex like the human mind is simple: [...] input, process, output. But our souls and what we do, how we live, everything inside it, that life and mind confuses things and complicates things because that's where the differences come.* [Alexander, AV-EXT, dropped out of therapy]

Significant Impact of TA

TA was a central aspect of therapy, with 14 participants noting its importance for experience, engagement, and outcomes. Trusting the therapist was essential for participants to remain engaged despite the emotional challenges of working with verbatim voice content.

> *I was able to trust the process and trust her, but I had to trust her to trust the process.* [Paula, AV-BRF, completed therapy]

Strong TA could also mitigate distressing aspects of therapy, such as voices responding negatively to progress in dialogues.

> *The voices themselves [...] were like dismissive of it [...] but when you have someone that [...] understands you, it does make you feel better, and the voices don't-can't change that.* [Joshua, AV-BRF, completed therapy]

While not identified as a notable issue for most, 2 participants described difficult feelings in relation to knowing that it was the therapist, someone they trusted, voicing derogatory voice content and reenacting abuse within avatar dialogues, particularly given the level of personal importance attached to the therapeutic relationship. Additionally, 4 participants highlighted that TA facilitated a clear separation between the avatar and the therapist.

> *Sometimes, knowing it's her that's saying it [...] it hurt a bit [laughs] because I liked her.* [Charlotte, AV-BRF, completed therapy]

The experience of TA itself had a positive impact in the short- and long-term, enhancing confidence and self-compassion and reducing stigma and self-blame.

> *The experience was empowering for me because I felt [...] I was not a patient being given treatment, I feel like I was treated as an equal.* [Stephen, AV-BRF, completed therapy]

This positive interpersonal experience reduced loneliness and improved trust in other relationships.

> *Beginning to gain trust again, it was slow, but it definitely opened the door to trusting again.* [Paula, AV-BRF, completed therapy]

Even those who did not complete therapy found this relationship impactful, and the therapy completers who did not perceive notable benefits (Zahid, Georgia, and Mai Su) found value in speaking openly with their therapist despite feeling that the AVATAR approach was not right for them at the time, demonstrating the importance of TA. For example, Alexander had previously lost hope that things could get better for him, as he felt no one cared, so this experience presented a significant catalyst for change.

> *People being genuinely supportive and actually trying to help people like me was enough for me. It felt like, no, maybe people do actually care, which was one of the reasons why I gave up in the first place.* [Alexander, AV-EXT, dropped out of therapy]

Theme 3: Individual Approach and Experience

This theme emphasizes the impact of individual participants' approaches and experiences. The subthemes "Open-minded attitudes despite initial doubts" and "Determination facilitated engagement and outcomes" highlight personal strengths. The subtheme "Profound emotional experience" focuses on the depth of emotional experience. The subtheme "Offered novel approach to tackle voices" considers the novelty of the therapy approach to the individual.

Open-Minded Attitudes Despite Initial Doubts

Most participants entered AVATAR therapy with some degree of anxiety, ranging from mild worry to skepticism.

> *Nervous [...] a bit scared.* [Georgia, AV-BRF, completed therapy]

Concerns included doubts about AVATAR therapy's effectiveness, the computerized approach, and fears that working directly with verbatim voice content might trigger negative psychological experiences.

> *I was very hesitant and reluctant to do it because, due to the nature of my voices, I had spent an extensive amount of time ignoring them and not engaging with them [...] so there was a fear of being like triggered or traumatised.* [Stephen, AV-BRF, completed therapy]

A total of 17 participants highlighted open-mindedness as central to initial engagement. Attitudes varied, and hope was often held cautiously; however, most were motivated to "give it a go," often driven by past challenges in accessing psychological support.

> *I just thought it doesn't hurt to give it a try.* [Zahid, AV-BRF, completed therapy]

Although 10 participants emphasized the central role of internal motivations, support from friends and family and trust in referring clinicians could enhance willingness to engage. Conversely, James described feeling pushed by his mental health team to participate despite his own reservations, which was disempowering and negatively impacted his engagement.

> *I wasn't really given a choice. I was told that I would want to do it, so therefore I should. And so I was just*

> *kind of shunted onto it.* [James, AV-EXT, dropped out of therapy]

Determination Facilitated Engagement and Outcomes

Determination was identified as important for maintaining engagement and facilitating positive outcomes for 17 participants. Motivations typically centered on a drive to overcome distressing voices and make progress, even when therapy processes felt uncomfortable.

> *It was just because I had a strong motivation to get past it all.* [Asim, AV-BRF, completed therapy]

Noticing gradual improvements could feel meaningful and reinforce motivations to continue. Paradoxically, Stephen found that noticing symptoms initially worsen actually increased his resolve to persevere.

> *The intensity of that after the first session was [...] trying to dissuade me, but I'd already made my mind up that I was going to do this and the fact that it did happen [...] gave me more motivation to continue to lean into it more.* [Stephen, AV-BRF, completed therapy]

Determination went beyond physically attending sessions, as 9 participants emphasized proactive efforts to be vulnerable and fully participate. In this way, the therapy experience and outcomes were perceived to be contingent on the depth of engagement and commitment to processes involved.

> *I think it varies what you put into it [...] that's what makes a big difference.* [William, AV-EXT, completed therapy]

However, engagement could be negatively impacted by external life stressors, which disrupted regular session attendance and motivations to keep going.

> *I'm glad I didn't have to do the 12-week sessions because that would have been quite hard just to fit in with all my [...] problems I had.* [Mai Su, AV-BRF, completed therapy]

Profound Emotional Experience

AVATAR therapy elicited a range of emotional experiences, which could be complex and difficult to articulate. Therapy was described as a meaningful journey by 15 participants, and the experience often felt transformative in that completing therapy led to personal growth, insights, and even a "paradigm shift."

> *It felt like, you know when people go on TV shows and they do social experiments, and then it's like this has changed my life like exponentially? It felt like I've been through this weird experience that I wouldn't change for the world.* [Stephen, AV-BRF, completed therapy]

Relevant to this meaningful journey, the emotional demands of therapy also presented challenges for 13 participants, such as reliving traumatic memories, confronting underlying problems, and feeling overwhelmed.

> *Difficult feelings of reliving what had happened to me [...] what I had put to the back of my mind was all coming out.* [Grace, AV-EXT, dropped out of therapy]

A total of 8 participants felt the depth of emotional experience enhanced outcomes, perceived as a challenging but necessary aspect of directly working through core issues. In this way, while experiences often became initially and temporarily more difficult, this eased across therapy and contributed to a meaningful sense of achievement when emerging stronger.

> *I really did feel like I was going back to the worst place [...] that I've been in terms of voices [...] that was a bit difficult but it passed [...] and I'm glad that it happened because I know how to deal with that better now.* [Gabriel, AV-EXT, completed therapy]

Additionally, therapy ending brought up mixed feelings. Some participants described a sense of closure and readiness to continue their recovery independently, while others experienced sadness, loss, and difficulties adjusting.

> *I felt complete.* [Stephen, AV-BRF, completed therapy]

> *I was sad, goodbyes are always hard for me.* [Caleb, AV-EXT, completed therapy]

Offered Novel Approach to Tackle Voices

The unique approach of working directly with voice content within avatar dialogues was highlighted as central to the success of therapy by 7 participants. Many valued the opportunity to experiment with different tactics, learn new skills, and practice responding to voices in real time, which was often a novel experience.

> *It gave me new perspectives on how to approach managing voices, like different ways to test out what would work for me.* [Gabriel, AV-EXT, completed therapy]

Some had never considered directly responding to voices, while others found avatar dialogues provided the opportunity to experience voices differently and achieve more effective communication. The avatar diverging from voices and conceding power could open new conversations and lead to voices also shifting. In this way, 11 participants valued the novel approach AVATAR therapy offered.

> *I would never have been able to have had this type of level of connection with the voice because my personal experience of it would never give that response.* [Stephen, AV-BRF, completed therapy]

However, Zahid felt AVATAR therapy offered him nothing significantly new, and although the avatar conceded power, this did not have any significant impact on voices.

> *It wasn't anything different that I haven't done.* [Zahid, AV-BRF, completed therapy]

Theme 4: Barriers to Engagement and Outcomes

The subthemes "Emotional challenges with avatar" and "Not the right approach for the individual at that time" focus on barriers and challenges faced by participants who dropped out of therapy, also relevant to some therapy completers. The subtheme "Difficulties translating changes to voices" relates to participants who completed therapy but struggled to access positive outcomes.

Emotional Challenges With Avatar

While most participants described early difficulties that eased over time (as discussed in "Initial challenges adjusting to avatar"), for some these were experienced as more significant emotional challenges. Such difficulties with digital representation could be challenging for some therapy completers to navigate, even among those who benefited, leading them to initially question their ability to continue.

> *Initially I found the avatar very difficult [...] I questioned whether I actually wanted to go through with it.* [Matthew, AV-EXT, completed therapy]

For the 3 completers who did not perceive benefit (Zahid, Georgia, and Mai Su), the avatar continued to feel either disconnected or frightening.

> *The avatar got a bit soft and started [...] back down a bit so it wasn't quite true to life, so I don't know.* [Mai Su, AV-BRF, completed therapy]

> *It was very deep.* [Georgia, AV-BRF, completed therapy]

All 3 participants who discontinued therapy (Alexander, James, and Grace) encountered emotional difficulties when working with a digital representation. Alexander reported feeling emotionally disconnected while cocreating the avatar, so he chose to drop out of therapy prior to any dialogues.

> *They are not the same. I was talking to a stranger. [...] There was no connection there at all.* [Alexander, AV-EXT, dropped out of therapy]

Although James also did not experience any avatar dialogues before deciding to discontinue therapy, he described the "crude" representation as an impassable barrier while acknowledging that the concept behind AVATAR therapy may be useful for some.

> *Freaked me out talking to this computer-generated head on the screen. It didn't look like anything that I [pause] knew and so I just [...] left because I couldn't engage in therapy with an avatar.* [James, AV-EXT, dropped out of therapy]

Grace found the avatar distressing and struggled with the emotional intensity. While she found the avatar a realistic representation and was able to meaningfully access therapy processes, she was fatigued by the sustained emotional intensity of dialogues and associated memories. This led her to drop out after 8 sessions, which involved 5 avatar dialogues.

> *I just felt I couldn't talk any longer because it was just too deep and I just I was getting home at night and I was just crying and I thought I just can't cope with this.* [Grace, AV-EXT, dropped out of therapy]

Not the Right Approach for the Individual at That Time

It was acknowledged that AVATAR therapy may not be suitable for everyone, depending on mental state and stage in recovery. For Alexander, choosing to discontinue therapy was experienced as empowering. The act of seeking help and developing a trusting relationship with his therapist provided a catalyst for change, emphasizing his inner strengths and jumpstarting his own path toward recovery. At this point, he felt therapy would be an unhelpful distraction, particularly in the context of autism and challenges with social interaction.

> *I was looking for the help and, as it turns out, the help wasn't outside, it was inside.* [Alexander, AV-EXT, dropped out of therapy]

Grace wondered if AVATAR therapy may have been easier for her to navigate earlier in her recovery and expressed a desire to learn strategies for managing difficult emotions, although she gained meaningful benefits from the 8 sessions (5 dialogues) she attended.

> *I don't think they could have done anything else, it was just up to me, I had to see what was good for me, what was not good for me.* [Grace, AV-EXT, dropped out of therapy]

Some therapy completers (n=5) suggested that effectiveness may vary depending on factors such as duration of voice-hearing experiences and readiness for therapy demands. In the context of their own challenges in early dialogues, it was speculated that working directly with voice content may be particularly difficult for those with "severe voices" and those who cope using avoidance.

> *I think it might just be a bit too overwhelming depending on the stage of and the seriousness of the voices.* [Ishan, AV-EXT, completed therapy]

Therefore, the importance of informed choice was emphasized, particularly given James' experiences of feeling pressured by his referrer. Some mentioned that they were given insufficient information about what to expect, and while acknowledging it may be difficult to grasp what it will be like until directly experiencing it, they suggested that showing a brief video of an avatar dialogue could help to reduce uncertainty.

> *Although everything's been explained, the words didn't help. I had to actually do it to find out.* [Alexander, AV-EXT, dropped out of therapy]

Difficulties Translating Changes to Voices

A barrier experienced by the 3 therapy completers who did not report notable benefits related to issues translating improvements from avatar dialogues to day-to-day interactions with voices. Zahid, Georgia, and Mai Su were able to stand up to the avatar but did not experience improvements with voices. Mai Su described struggling with independent practice and reverting to default responses.

> *When I'm at home and the voice talks to me and it's loud and I just don't remember any of those sessions, so I just scream back it.* [Mai Su, AV-BRF, completed therapy]

A total of 3 therapy completers who reported benefit from AVATAR therapy also faced challenges in independently applying the approach to voices, experiencing initial discomfort and hesitancy.

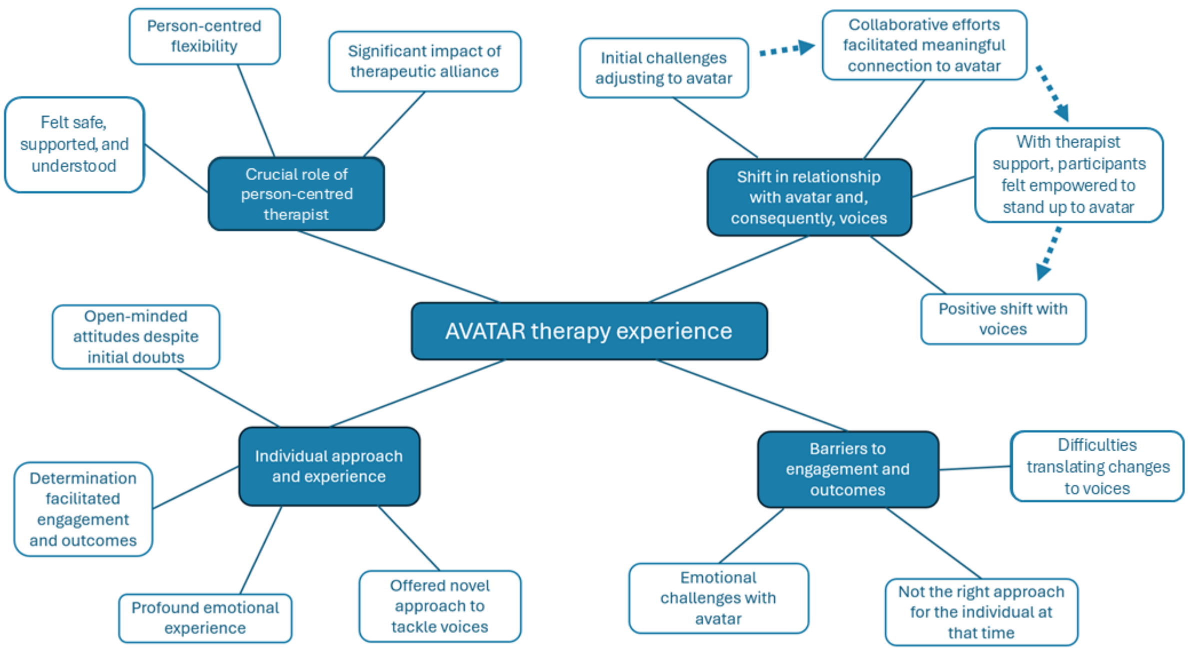
> *I didn't really practice too much [...] because I needed to build up more confidence [...] I felt like I wasn't ready yet in myself.* [Ishan, AV-EXT, completed therapy]

Progress was not always smooth or cumulative. For example, Asim found voices increased in aggression after completing therapy and valued an additional session with his therapist to regain control. This highlights that, even when AVATAR therapy is beneficial, there may be challenges in applying learning to daily life, and there could be benefits of booster sessions.

> *After I started using the AVATAR therapy, you know, the stuff I learned, it was okay for a while but it came back angrier.* [Asim, AV-BRF, completed therapy]

The network of themes and subthemes is summarized in Figure 1, which illustrates the 4 overarching themes and their interrelationships. The arrows between subthemes in the "Shift in relationship with avatar and, consequently, voices" theme highlight the sequential process of first adjusting to and building a connection with the avatar before standing up to it and gaining benefits.

Figure 1. Thematic network.



Discussion

Overview

This study used peer research methods and a dual approach of IPA and template analysis to explore the experience of AVATAR therapy, with a focus on direct work with voice content and the role of TA. Findings align with many of the themes identified in AVATAR1 qualitative work [11], particularly in highlighting the value of communicating with a digital representation, learning strategies, and the role of therapist support. Those who discontinued therapy described challenges with the realism of working dialogically with an avatar, reflecting that this approach may not be effective for everyone.

How Did Participants Experience Directly Working With Verbatim Voice Content?

Despite initial challenges with early exposure and in vivo communication, working dialogically with voice content led most to experience an increased sense of power and control over voices, consistent with other cognitive approaches for distressing voices [39]. Participant accounts aligned with core therapeutic aims and processes, particularly around empowerment and self-esteem [8,40]. Consistent with inhibitory learning theory [41], participants described challenging distressing beliefs and targeting safety behaviors (including submission) within avatar dialogues, resulting in reduced anxiety over time. Positive outcomes were varied and individualized, including reduced voice frequency and omnipotence, enhanced understanding and acceptance, and improved interpersonal functioning. This aligns with relational approaches for distressing voices [3,7,42] and reflects the broader secondary outcomes observed in the main trial, such as improved voice understanding and empowerment [9].

Although use of technology initially caused some anxiety, its role in externally representing voices was well accepted by most, mirroring AVATAR1 qualitative findings [11]. Participant accounts regarding their emotional connection to, or disconnection from, the avatar relate to the concept of sense of voice presence [12,13]. This is a key mechanism in fear activation [43], promoting emotional processing and generalization of learning [44,45]. In line with earlier work [13], most participants reported at least some degree of voice presence during avatar dialogues. However, early challenges involved either high levels of perceived realism leading to fear activation (particularly in the context of safety behaviors, especially avoidance, and reactivation of trauma memories) or difficulties accessing voice presence inhibiting emotional processes. This reflects the bidirectional relationship between presence and emotion [46]. Participant accounts support the assertion that engagement and positive outcomes are contingent on voices being brought online within dialogues to activate emotion (ie, anxiety) within a safe therapeutic space [13].

Some participants indicated that, in the context of a high sense of voice presence, they experienced temporary symptom exacerbation. However, consistent with previous findings [13], the emotional intensity of this experience appeared well-tolerated by most and promptly eased across therapy. In fact, many specifically valued coming through this experience because it meant they had directly confronted and overcome core issues. This is consistent with evidence of temporary symptom exacerbation when working through challenges in trauma-focused imaginal exposure [47] and other exposure therapies [48,49].

Participants identified how open-minded attitudes, resilience, and determination helped to overcome the challenge of direct work with voice content, particularly when navigating anxiety-provoking early dialogues. While often conflated with attendance [50], many emphasized that engagement went beyond simply turning up each week and involved "leaning into" therapy to access voice presence, emotional processes, and meaningful outcomes. The emotional experience of therapy ending was also highly individualized across both AV-BRF and AV-EXT, with some experiencing a profound sense of closure and others struggling with loss. This variability reflects individual differences and the diverse meanings and relationships involved in voice-hearing, highlighting that the experience of therapy ending can be shaped by personal meaning and relational factors [51].

Generalizing learning from avatar dialogues to voices in daily life could be challenging, reflecting broader literature and emphasizing the need for continued use of skills [52]. The 3 therapy completers did not experience notable positive outcomes, possibly reflecting reported challenges in accessing voice presence, as this has been linked to the transfer of knowledge and skills [13].

It is important to note that participants who discontinued therapy were underanalyzed in relation to voice presence, with only 1 included in the theme "Collaborative efforts facilitated meaningful connection to avatar." The 2 other participants disengaged prior to any dialogues in the context of challenges with the realism of the software, meaning they did not reach the point of establishing presence. These issues are further considered in relation to dropout experiences below.

How Was TA Experienced by Participants in the Context of AVATAR Therapy?

Participant accounts align with the concept of TA [53], emphasizing therapist qualities like warmth and empathy. Strong TA was established for all, including those who dropped out, underscoring its central role in therapy experience, engagement, and outcomes, consistent with other psychosis interventions [54-56]. Developing trust was perceived to be vital given sensitive therapy processes, such as sharing verbatim voice content, exposure to feared stimuli, and working with trauma. In particular, participants valued transparent, respectful conversations to ensure an appropriate balance of voice presence and safety, particularly when reenacting abuse via the avatar [8]. Therefore, TA could mitigate the impact of the therapist operating the avatar and facilitate continued engagement despite anxiety. Therapists' emotional attunement within avatar dialogues enabled person-centered therapeutic input, which progressively reduced across therapy [40].

Despite initial concerns about the impact of technology on TA, most participants felt they had adequate access to human therapist support. In fact, the digitally mediated environment was perceived to provide therapists with unique insights into participants' live experiences and behaviors, creating valuable opportunities for validation. Participants also valued therapists' respectful curiosity about their experiences, values, beliefs, strengths, and cultural background. In this way, many described feeling empowered to take an active role in their therapy, emphasizing collaboration and choice [57,58].

What Were the Experiences of Participants Who Decided Not to Continue With Therapy?

Choosing to discontinue therapy could be an empowering choice, providing insights for recovery, but could also be experienced negatively, reflecting the contrasting experiences of therapy dropout [59]. This also emphasizes that attendance is only one component of engagement, and it is possible to meaningfully access therapeutic processes without reaching a treatment dose [50].

While challenges of exposure to the avatar eased promptly for most, the emotional toll of dialogues across therapy led 1 participant to drop out after 5 dialogues (8 sessions). Additionally, 2 other participants chose to discontinue prior to any dialogues due to difficulties with the realism of the avatar software. This suggests that AVATAR therapy may be overwhelming if voice-hearing experiences elicit high levels of anxiety [60], and, conversely, insufficient realism may disrupt voice presence and engagement. In this way, both too much and too little realism can create difficulties in establishing voice presence and activating target emotions in a safe setting, posing a risk for disengagement. However, many therapy completers who benefited from AVATAR therapy reported similar challenges in early sessions and initially questioned whether to continue, highlighting the role of personal choice. Factors such as life events outside of therapy, access to resources, and social functioning may influence an individual's decision to continue with therapy despite challenges [61-63]. This was endorsed by therapy completers who acknowledged that AVATAR therapy may not be effective for all, depending on factors such as stage in recovery, stressful life events, and physical health, consistent with wider literature [47,64-66]. Taken together, this indicates that dropout may not simply reflect limitations in software realism but also readiness for exposure and suitability of the intervention for the individual at that time.

Similarly to AVATAR1 qualitative findings [11], 1 autistic participant struggled to engage with the idea of digitally embodying the voice, possibly suggesting additional challenges for those with this comorbidity. This highlights the need for further research specifically focused on autistic voice-hearers.

Strengths and Limitations

The study demonstrated methodological rigor with adherence to peer research methods and IPA principles, such as using a phenomenological approach, ensuring reflexivity, and using an iterative process of analysis [18]. Steps were taken to standardize interview delivery through training, role-play, co-interviewers, and supervision. The peer researcher status of interviewers likely supported participants to feel comfortable, understood, and willing to open up [67]. Additionally, efforts were made to include participants who discontinued AVATAR therapy, enhancing learning around potential reasons for making this decision.

While integrating 2 qualitative analysis methods enabled both depth and breadth [36], this introduced epistemological tension as IPA emphasizes idiographic depth while template analysis prioritizes cross-case patterns. The careful layering of these methods was guided by advice from experts in qualitative research methods and supported by reflexive practices and peer research methods. However, some complexity may have been reduced when moving to broader themes, with the risk of diluting the nuance of individual voices. This highlights the trade-off between depth and generalizability [21].

Selecting transcripts for IPA based on richness of data (often defined in terms of depth, detail, and context provided within narratives) has been questioned due to the associated risk of bias in privileging some accounts over others [22]. In line with IPA principles, transcripts of participants who discontinued therapy were not included in IPA to ensure homogeneity [19,34]. While the integration of template analysis enabled clear comparison of completers and noncompleters and ensured that all participants' perspectives were systematically incorporated [23], this may have privileged voices that were more positive and engaged. This could limit the transferability of findings, as the therapeutic processes identified in detail may not fully reflect the experiences of those who disengaged or found the intervention unsuitable.

Participants were invited for an interview after completing the 28-week follow-up to avoid any influence on trial outcomes. Therefore, interviews were conducted several months after completing AVATAR therapy, enabling exploration of long-term perceived impacts. However, this delay may have also introduced recall bias or retrospective reinterpretation of participants' experiences during therapy. Social desirability bias and self-selection bias may have impacted study recruitment. While we recruited a broadly representative sample with increased numbers of participants who dropped out of therapy compared to AVATAR1, only 4.3% (3/69) of those who dropped out were represented compared to 9.9% (16/161) of therapy completers. As it was not possible to recruit the intended 4 participants who dropped out, and the interview sample included 2 participants who dropped out prior to any dialogues, we have likely missed valuable experiences and perspectives from those who were not willing to participate. Additionally, as this study focused on TA between the voice-hearer and therapist, we may have missed important insights into the "triangle of alliance" between the participant, therapist, and digital platform (ie, avatar) [68].

A total of 10.5% (2/19) of participants were from Black backgrounds compared to 17.8% (41/230) in the wider AVATAR2 sample. The experiences of Black participants are the focus of a stand-alone qualitative study within the AVATAR2 Programme of Qualitative Research, which is currently being prepared for publication [69]. Concomitant recruitment for both qualitative studies is a likely factor in the lower representation of Black participants within this study.

Implications for Optimization and Implementation

The AVATAR2 trial aimed to evaluate the efficacy of 2 forms of AVATAR therapy, determine optimal therapy delivery, and consider implementation in National Health Service (NHS) settings. Supported by a National Institute for Health and Care Excellence Early Value Assessment [70], which has recently recommended further testing in routine NHS settings, these qualitative findings will support the optimization and wider implementation of this approach.

Establishing voice presence and activating target emotions within a safe setting has been emphasized as central to engagement and outcomes. Software developments may improve the realism of digital representation, enhancing sense of voice presence. Findings emphasize the need for improvements to the tailoring of avatars across different ethnicities and genders to ensure equal opportunities to establish voice presence. However, as technology advances, increased realism might activate strong emotional responses and increase fear within the digital environment, particularly pertinent given the ongoing trial of AVATAR-VR [71]. Therefore, flexible therapist support will be crucial to ensure participants feel safe, balancing exposure with tolerability [8,47]. Meanwhile, dropout experiences suggest that difficulties establishing presence may not only reflect technological limitations in software realism but also individual readiness and suitability, emphasizing the importance of personalizing AVATAR therapy to the person's stage of recovery and capacity to tolerate exposure.

Limitations suggest that our findings are most transferable for understanding therapeutic processes when engagement is maintained, providing less insight into disengagement and nonresponse. Therefore, future work should further explore the perspectives of those who discontinue or gain limited benefit. Such insights could inform early screening and tailoring, for example, around readiness for exposure and tolerance of voice presence.

In this context, integrating emotion regulation and anxiety management techniques may enhance coping and sense of safety [47,72], as demonstrated by Paulik et al [73] in supporting voice-hearers to engage with exposure therapies. Therefore, future research could consider if an emotion regulation phase improves the tolerance of this experience, ensuring trauma-informed care and integrating other therapeutic approaches [47]. Additionally, tools to support normalizing discussions around common voice content and transparency over potential temporary voice exacerbation could facilitate collaborative engagement and reduce distress.

Findings highlight the importance of TA, fully informed consent, and patient choice when proceeding with exposure and dialogues. Peer support spaces could provide additional support, increase motivation, and promote engagement and outcomes [74,75]. Further research could explore the triadic relationship between the participant, therapist, and avatar. In particular, it would be valuable to examine how the concept of TA applies to the avatar, given its role in representing the distressing voice yet also the individualized relational context, meaning that there may be a therapeutic focus on fostering compassion toward the voice [8,76].

The range of emotional experiences across AV-BRF and AV-EXT highlights the complexities and relevance of individual differences. AVATAR2 RCT findings suggest that AV-EXT produces stronger results, but this was in the context of higher dropout rates. As AVATAR therapy moves from clinical trials and into routine care, there are increased opportunities for tailoring the therapy to context, including flexibility in duration, use of booster sessions, and potential integration into long-term therapy.

Conclusions

This study provides valuable insights into the participant experience of AVATAR therapy, enriched by peer research methods. Overall, experiences were positive, with all reporting strong TA and most gaining power over the avatar and, consequently, voices. There were challenges with in vivo exposure and dialoguing with voice content, but this was often mitigated by TA. Engagement and positive outcomes were contingent on establishing voice presence and activating target emotions while retaining a sense of safety. Potential barriers to engagement and outcome have been suggested, emphasizing the need for software developments, person-centered flexibility, and informed choice.

Acknowledgments

We would like to thank the following people who contributed to this study but are not listed as authors. First and foremost, we thank all the participants who consented to take part and generously shared their experiences. We would also like to thank the patient and public involvement consultants for their valuable contributions at every stage of this study and the following research workers for their input in facilitating the peer research approach: Leonie Richardson, Dan Ramsell, Tami Sowemimo, Rupa Patel, Alicia McGee, Oliver Owrid, Stephanie Allan, and Lucy Miller. We also thank Dan Robotham, Michael Larkin, and Elena Gil-Rodriguez for sharing their expertise in qualitative research methods and contributing to the analytic approach. We would like to highlight that Dan Robotham's contributions were made during his role at the McPin Foundation, prior to joining Wellcome.

Funding

This research was funded by a Wellcome grant 215471/Z/19/Z. TW acknowledges funding from the Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. ERE acknowledges support for this project through the Doctorate in Clinical Psychology training program at King's College London, which provided additional funding. The study was supported by the National Institute for Health and Care Research (NIHR), the NIHR Maudsley Biomedical Research Centre and Maudsley NHS Foundation Trust, King's College London (PG; NIHR203318), and the Manchester Biomedical Research Centre (SB and GH; NIHR203308). SB discloses support for publication of this work from the NIHR research professorship (NIHR300794). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. This study was supported by NHS Research Scotland through the Chief Scientist Office and the NHS Scotland Mental Health Network (A Gumley; NRSMHN/2021/01). MRC discloses support for the publication of this work from the Sofja Kovalevskaja Award from the Alexander von Humbold Foundation and the Ministry of Education and Research in Germany (SKP2020RusCalafell). The funding bodies did not have any role in study design; data collection, analysis, and interpretation; or writing of the manuscript.

Data Availability

This study was preregistered in the Open Science Framework [30]. Although anonymized, the qualitative data generated for this study are potentially identifiable, and participants only provided written informed consent for anonymized sections to be published. Therefore, the qualitative datasets are not suitable to be deposited in a public database. However, the data are available from the corresponding author on reasonable request.

Authors' Contributions

Study design: TW, CE, MRC, A Gumley, ERE, PG, TC
Data collection (interviews): ERE, NB, AP, NH
Data analysis (interpretative phenomenological analysis): ERE with input from NB, AP, NH
Coding template development and theme refinement: ERE, TW, CE, NB, JH, AP, CMN, NH, A Grant
Manuscript drafting: ERE

Critical manuscript review: TW, CE, TC, PG, MRC, HB, MC, JM, A Gumley, GH, SB, MFA
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Stages of PPI.
[PDF File (Adobe PDF File), 22 KB - mental_v13i1e77566_app1.pdf]

Multimedia Appendix 2
Interview guide.
[DOCX File , 28 KB - mental_v13i1e77566_app2.docx]

Multimedia Appendix 3
Interview training materials.
[PDF File (Adobe PDF File), 366 KB - mental_v13i1e77566_app3.pdf]

Multimedia Appendix 4
Excerpts of analytic processes.
[DOCX File , 4113 KB - mental_v13i1e77566_app4.docx]

Multimedia Appendix 5
Provisional template.
[DOCX File , 127 KB - mental_v13i1e77566_app5.docx]

Multimedia Appendix 6
Summary of themes.
[DOCX File , 20 KB - mental_v13i1e77566_app6.docx]

Multimedia Appendix 7
Additional quotes.
[DOCX File , 32 KB - mental_v13i1e77566_app7.docx]

References

1. Linscott R, van Os J. An updated and conservative systematic review and meta-analysis of epidemiological evidence on psychotic experiences in children and adults: on the pathway from proneness to persistence to dimensional expression across mental disorders. Psychol Med 2012;43(6):1133-1149. [doi: 10.1017/S0033291712001626] [Medline: 22850401]
2. Sorrell E, Hayward M, Meddings S. Interpersonal processes and hearing voices: a study of the association between relating to voices and distress in clinical and non-clinical hearers. Behav Cogn Psychother 2010;38(2):127-140. [doi: 10.1017/S1352465809990506] [Medline: 19878609]
3. Steel C, Schnackenberg J, Perry H, Longden E, Greenfield E, Corstens D. Making sense of voices: a case series. Psychosis 2019;11(1):3-15 [FREE Full text] [doi: 10.1080/17522439.2018.1559874]
4. Corstens D, Longden E, May R. Talking with voices: exploring what is expressed by the voices people hear. Psychosis 2012;4(2):95-104 [FREE Full text] [doi: 10.1080/17522439.2011.571705]
5. Hayward M, Jones A, Bogen-Johnston L, Thomas N, Strauss C. Relating therapy for distressing auditory hallucinations: a pilot randomized controlled trial. Schizophr Res 2017;183:137-142. [doi: 10.1016/j.schres.2016.11.019] [Medline: 27916286]
6. Leff J, Williams G, Huckvale MA, Arbuthnot M, Leff AP. Computer-assisted therapy for medication-resistant auditory hallucinations: proof-of-concept study. Br J Psychiatry 2013;202:428-433. [doi: 10.1192/bjp.bp.112.124883] [Medline: 23429202]
7. Craig T, Rus-Calafell M, Ward T, Leff J, Huckvale M, Howarth E, et al. AVATAR therapy for auditory verbal hallucinations in people with psychosis: a single-blind, randomised controlled trial. Lancet Psychiatry 2018;5(1):31-40 [FREE Full text] [doi: 10.1016/S2215-0366(17)30427-3] [Medline: 29175276]
8. Ward T, Rus-Calafell M, Ramadhan Z, Soumelidou O, Fornells-Ambrojo M, Garety P, et al. AVATAR therapy for distressing voices: a comprehensive account of therapeutic targets. Schizophr Bull 2020;46(5):1038-1044 [FREE Full text] [doi: 10.1093/schbul/sbaa061] [Medline: 32372082]

9. Garety P, Edwards C, Jafari H, Emsley R, Huckvale M, Rus-Calafell M, et al. Digital AVATAR therapy for distressing voices in psychosis: the phase 2/3 AVATAR2 trial. Nat Med 2024;30(12):3658-3668. [doi: 10.1038/s41591-024-03252-8] [Medline: 39468363]
10. Longden E, Corstens D, Escher S, Romme M. Voice hearing in a biographical context: a model for formulating the relationship between voices and life history. Psychosis 2012;4(3):224-234 [FREE Full text] [doi: 10.1080/17522439.2011.596566]
11. Rus-Calafell M, Ehrbar N, Ward T, Edwards C, Huckvale M, Walke J, et al. Participants' experiences of AVATAR therapy for distressing voices: a thematic qualitative evaluation. BMC Psychiatry 2022;22(1):356 [FREE Full text] [doi: 10.1186/s12888-022-04010-1] [Medline: 35610590]
12. Slater M. A note on presence terminology. Presence Connect 2003;3(3):1-5 [FREE Full text]
13. Rus-Calafell M, Ward T, Zhang X, Edwards C, Garety P, Craig T. The role of sense of voice presence and anxiety reduction in AVATAR therapy. J Clin Med 2020;9(9):2748 [FREE Full text] [doi: 10.3390/jcm9092748] [Medline: 32854387]
14. Henson P, Wisniewski H, Hollis C, Keshavan M, Torous J. Digital mental health apps and the therapeutic alliance: initial review. BJPsych Open 2019;5(1):e15 [FREE Full text] [doi: 10.1192/bjo.2018.86] [Medline: 30762511]
15. Tremain H, McEnery C, Fletcher K, Murray G. The therapeutic alliance in digital mental health interventions for serious mental illnesses: narrative review. JMIR Ment Health 2020;7(8):e17204 [FREE Full text] [doi: 10.2196/17204] [Medline: 32763881]
16. Tong F, Lederman R, D'Alfonso S, Berry K, Bucci S. Conceptualizing the digital therapeutic alliance in the context of fully automated mental health apps: a thematic analysis. Clin Psychol Psychother 2023;30(5):998-1012. [doi: 10.1002/cpp.2851] [Medline: 37042076]
17. Brotherdale R, Berry K, Bucci S. A qualitative study exploring the digital therapeutic alliance with fully automated smartphone apps. Digit Health 2024;10:20552076241277712 [FREE Full text] [doi: 10.1177/20552076241277712] [Medline: 39687527]
18. Smith JA, Flowers P, Larkin M. Interpretative Phenomenological Analysis: Theory, Method and Research. London: Sage; 2009.
19. Smith JA, Flowers P, Larkin M. Interpretative Phenomenological Analysis: Theory, Method, Research. London: Sage; 2022.
20. Smith J, Osborn M. Interpretative phenomenological analysis as a useful methodology for research on the lived experience of pain. Br J Pain 2015;9(1):41-42 [FREE Full text] [doi: 10.1177/2049463714541642] [Medline: 26516556]
21. King N. Doing template analysis. In: Symon G, Cassell C, editors. Qualitative Organizational Research: Core Methods and Current Challenges. London: Sage; 2012:426-450.
22. Frost N. Interpreting data pluralistically. In: Qualitative Research Methods in Psychology: Combining Core Approaches. London, Greater London: Open University Press; 2011:145-160.
23. Brooks J, McCluskey S, Turley E, King N. The utility of template analysis in qualitative psychology research. Qual Res Psychol 2015;12(2):202-222 [FREE Full text] [doi: 10.1080/14780887.2014.955224] [Medline: 27499705]
24. Bond J, Kenny A, Pinfold V, Couperthwaite L, gameChange Lived Experience Advisory Panel, Kabir T, et al. A safe place to learn: peer research qualitative investigation of gameChange virtual reality therapy. JMIR Serious Games 2023;11:e38065 [FREE Full text] [doi: 10.2196/38065] [Medline: 36645707]
25. Smith JA, Jarman M, Osborn M. Doing interpretative phenomenological analysis. In: Qualitative Health Psychology: Theories and Methods. Thousand Oaks, CA: SAGE Publications; 1999:218-240.
26. Sweeney A, Greenwood K, Williams S, Wykes T, Rose D. Hearing the voices of service user researchers in collaborative qualitative data analysis: the case for multiple coding. Health Expect 2013;16(4):e89-e99 [FREE Full text] [doi: 10.1111/j.1369-7625.2012.00810.x] [Medline: 22958162]
27. Gillard S, Borschmann R, Turner K, Goodrich-Purnell N, Lovell K, Chambers M. 'What difference does it make?' Finding evidence of the impact of mental health service user researchers on research into the experiences of detained psychiatric patients. Health Expect 2010;13(2):185-194 [FREE Full text] [doi: 10.1111/j.1369-7625.2010.00596.x] [Medline: 20536538]
28. Haven T, Errington T, Gleditsch K, van Grootel L, Jacobs A, Kern F, et al. Preregistering qualitative research: a delphi study. Int J Qual Method 2020;19 [FREE Full text] [doi: 10.1177/1609406920976417]
29. Staniszewska S, Brett J, Simera I, Seers K, Mockford C, Goodlad S, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. BMJ 2017;358:j3453 [FREE Full text] [doi: 10.1136/bmj.j3453] [Medline: 28768629]
30. AVATAR2 research programme: work package 1 - AVATAR therapy for distressing voices: exploring the therapy experience. Open Science Framework. 2023. URL: https://osf.io/h87cq [accessed 2025-12-04]
31. UK standards for public involvement in research. INVOLVE. 2019. URL: https://www.invo.org.uk/wp-content/uploads/2019/11/UK-standards-for-public-involvement-v6.pdf [accessed 2025-12-03]
32. Owrid O, Richardson L, Allan S, Grant A, Gogan S, Hamilton N, et al. "There's no us vs. them, it's just us": a creative approach to centring lived experience within the AVATAR2 trial. BMC Psychiatry 2024;24(1):807 [FREE Full text] [doi: 10.1186/s12888-024-06268-z] [Medline: 39543539]

33. Garety P, Edwards C, Ward T, Emsley R, Huckvale M, McCrone P, et al. Optimising AVATAR therapy for people who hear distressing voices: study protocol for the AVATAR2 multi-centre randomised controlled trial. Trials 2021;22(1):366 [FREE Full text] [doi: 10.1186/s13063-021-05301-w] [Medline: 34034792]
34. Larkin M, Shaw R, Flowers P. Multiperspectival designs and processes in interpretative phenomenological analysis research. Qual Res Psychol 2018;16(2):182-198 [FREE Full text] [doi: 10.1080/14780887.2018.1540655]
35. Harding R, Whitfield G, Stillwell N. Service users as peer research interviewers: why bother? In: Social Policy Review 22: Analysis and Debate in Social Policy. Bristol UK: Policy Press; 2010.
36. Bond J, Robotham D, Kenny A, Pinfold V, Kabir T, Andleeb H, et al. Automated virtual reality cognitive therapy for people with psychosis: protocol for a qualitative investigation using peer research methods. JMIR Res Protoc 2021;10(10):e31742 [FREE Full text] [doi: 10.2196/31742] [Medline: 34694236]
37. Smith J, Nizza I. Essentials of Interpretative Phenomenological Analysis. Washington, D.C: American Psychological Association; 2022.
38. Jacobson D, Mustafa N. Social identity map: a reflexivity tool for practicing explicit positionality in critical qualitative research. Int J Qual Method 2019;18 [FREE Full text] [doi: 10.1177/1609406919870075]
39. Birchwood M, Dunn G, Meaden A, Tarrier N, Lewis S, Wykes T, et al. The COMMAND trial of cognitive therapy to prevent harmful compliance with command hallucinations: predictors of outcome and mediators of change. Psychol Med 2018;48(12):1966-1974 [FREE Full text] [doi: 10.1017/S0033291717003488] [Medline: 29202885]
40. O'Brien C, Rus-Calafell M, Craig T, Garety P, Ward T, Lister R, et al. Relating behaviours and therapeutic actions during AVATAR therapy dialogue: an observational study. Br J Clin Psychol 2021;60(4):443-462. [doi: 10.1111/bjc.12296] [Medline: 33949726]
41. Craske M, Treanor M, Conway C, Zbozinek T, Vervliet B. Maximizing exposure therapy: an inhibitory learning approach. Behav Res Ther 2014;58:10-23 [FREE Full text] [doi: 10.1016/j.brat.2014.04.006] [Medline: 24864005]
42. Longden E, Corstens D, Bowe S, Pyle M, Emsley R, Peters S, et al. A psychological intervention for engaging dialogically with auditory hallucinations (Talking With Voices): a single-site, randomised controlled feasibility trial. Schizophr Res 2022;250:172-179 [FREE Full text] [doi: 10.1016/j.schres.2022.11.007] [Medline: 36423442]
43. Wiederhold BK, Wiederhold MD. The effect of presence on virtual reality treatment. In: Virtual Reality Therapy for Anxiety Disorders: Advances in Evaluation and Treatment. Washington, DC: American Psychological Association; 2005:77-86.
44. Ling Y, Nefs H, Morina N, Heynderickx I, Brinkman W. A meta-analysis on the relationship between self-reported presence and anxiety in virtual reality exposure therapy for anxiety disorders. PLoS One 2014;9(5):e96144 [FREE Full text] [doi: 10.1371/journal.pone.0096144] [Medline: 24801324]
45. Foa E, Kozak M. Emotional processing of fear: exposure to corrective information. Psychological Bulletin 1986;99(1):20-35 [FREE Full text] [doi: 10.1037/0033-2909.99.1.20]
46. Riva G, Mantovani F. Being there: understanding the feeling of presence in a synthetic environment and its potential for clinical change. In: Januszewski M, editor. Virtual Reality in Psychological, Medical and Pedagogical Applications. London: InTech Open; 2012:3-4.
47. Feary N, Brand R, Williams A, Thomas N. 'Like jumping off a ledge into the water': a qualitative study of trauma-focussed imaginal exposure for hearing voices. Psychol Psychother 2022;95(1):277-294 [FREE Full text] [doi: 10.1111/papt.12372] [Medline: 34799984]
48. Tong J, Simpson K, Alvarez-Jimenez M, Bendall S. Distress, psychotic symptom exacerbation, and relief in reaction to talking about trauma in the context of beneficial trauma therapy: perspectives from young people with post-traumatic stress disorder and first episode psychosis. Behav Cogn Psychother 2017;45(6):561-576. [doi: 10.1017/S1352465817000236] [Medline: 28436349]
49. Burger S, Hardy A, van der Linden T, van Zelst C, de Bont PAJ, van der Vleugel B, et al. The bumpy road of trauma-focused treatment: posttraumatic stress disorder symptom exacerbation in people with psychosis. J Trauma Stress 2023;36(2):299-309. [doi: 10.1002/jts.22907] [Medline: 36719408]
50. Holdsworth E, Bowen E, Brown S, Howat D. Client engagement in psychotherapeutic treatment and associations with client characteristics, therapist characteristics, and treatment factors. Clin Psychol Rev 2014;34(5):428-450. [doi: 10.1016/j.cpr.2014.06.004] [Medline: 25000204]
51. Råbu M, Binder P, Haavind H. Negotiating ending: a qualitative study of the process of ending psychotherapy. Eur J Psychother Couns 2013;15(3):274-295 [FREE Full text] [doi: 10.1080/13642537.2013.810962]
52. Hoet AC, Burgin CJ, Eddington KM, Silvia PJ. Reports of therapy skill use and their efficacy in daily life in the short-term treatment of depression. Cogn Ther Res 2017;42(2):184-192 [FREE Full text] [doi: 10.1007/s10608-017-9852-y]
53. Bordin E. The generalizability of the psychoanalytic concept of the working alliance. Psychol Psychother Theory Res Prac 1979;16(3):252-260 [FREE Full text] [doi: 10.1037/h0085885]
54. Browne J, Nagendra A, Kurtz M, Berry K, Penn D. The relationship between the therapeutic alliance and client variables in individual treatment for schizophrenia spectrum disorders and early psychosis: narrative review. Clin Psychol Rev 2019;71:51-62. [doi: 10.1016/j.cpr.2019.05.002] [Medline: 31146249]

55. Bourke E, Barker C, Fornells-Ambrojo M. Systematic review and meta-analysis of therapeutic alliance, engagement, and outcome in psychological therapies for psychosis. Psychol Psychother 2021;94(3):822-853. [doi: 10.1111/papt.12330] [Medline: 33569885]
56. Goldsmith L, Lewis S, Dunn G, Bentall R. Psychological treatments for early psychosis can be beneficial or harmful, depending on the therapeutic alliance: an instrumental variable analysis. Psychol Med 2015;45(11):2365-2373 [FREE Full text] [doi: 10.1017/S003329171500032X] [Medline: 25805118]
57. Priebe S, Watts J, Chase M, Matanov A. Processes of disengagement and engagement in assertive outreach patients: qualitative study. Br J Psychiatry 2005;187:438-443. [doi: 10.1192/bjp.187.5.438] [Medline: 16260819]
58. Williams A, Fossey E, Farhall J, Foley F, Thomas N. Impact of jointly using an e-mental health resource (self-management and recovery technology) on interactions between service users experiencing severe mental illness and community mental health workers: grounded theory study. JMIR Ment Health 2021;8(6):e25998 [FREE Full text] [doi: 10.2196/25998] [Medline: 34132647]
59. Simon G, Imel Z, Ludman E, Steinfeld B. Is dropout after a first psychotherapy visit always a bad outcome? Psychiatr Serv 2012;63(7):705-707 [FREE Full text] [doi: 10.1176/appi.ps.201100309] [Medline: 22752034]
60. Simpson D, Rowan-Szal G, Joe G, Best D, Day E, Campbell A. Relating counselor attributes to client engagement in England. J Subst Abuse Treat 2009;36(3):313-320 [FREE Full text] [doi: 10.1016/j.jsat.2008.07.003] [Medline: 18835675]
61. Reiter MD. Hope and expectancy in solution-focused brief therapy. J Fam Psychother 2010;21(2):132-148 [FREE Full text] [doi: 10.1080/08975353.2010.483653]
62. Duncan B, Miller S. The client's theory of change: consulting the client in the integrative process. J Psychother Integr 2000;10:169-187 [FREE Full text] [doi: 10.1023/A:1009448200244]
63. Lincoln T, Rief W, Westermann S, Ziegler M, Kesting M, Heibach E, et al. Who stays, who benefits? Predicting dropout and change in cognitive behaviour therapy for psychosis. Psychiatry Res 2014;216(2):198-205 [FREE Full text] [doi: 10.1016/j.psychres.2014.02.012] [Medline: 24602992]
64. Bados A, Balaguer G, Saldaña C. The efficacy of cognitive-behavioral therapy and the problem of drop-out. J Clin Psychol 2007;63(6):585-592. [doi: 10.1002/jclp.20368] [Medline: 17457848]
65. Bourdeau G, Lecomte T, Lysaker P. Stages of recovery in early psychosis: associations with symptoms, function, and narrative development. Psychol Psychother 2015;88(2):127-142. [doi: 10.1111/papt.12038] [Medline: 25139504]
66. Scott J, Leboyer M, Hickie I, Berk M, Kapczinski F, Frank E, et al. Clinical staging in psychiatry: a cross-cutting model of diagnosis with heuristic and practical value. Br J Psychiatry 2013;202(4):243-245. [doi: 10.1192/bjp.bp.112.110858] [Medline: 23549937]
67. Simpson E, House A. Involving users in the delivery and evaluation of mental health services: systematic review. BMJ 2002;325(7375):1265 [FREE Full text] [doi: 10.1136/bmj.325.7375.1265] [Medline: 12458241]
68. Cavanagh K. Turn on, tune in and (don?t) drop outngagement, adherence, attrition, and alliance with internet-based interventions. In: Oxford Guide to Low Intensity CBT Interventions. Oxford: Oxford University Press; 2010:227-234.
69. Black people's experience of AVATAR therapy. OSF Project. 2023. URL: https://osf.io/qrykz [accessed 2025-12-03]
70. Early value assessment interim statement [PMG39]. NICE. 2022. URL: https://www.nice.org.uk/process/pmg39 [accessed 2025-12-03]
71. Smith L, Mariegaard L, Vernal D, Christensen A, Albert N, Thomas N, et al. The CHALLENGE trial: the effects of a virtual reality-assisted exposure therapy for persistent auditory hallucinations versus supportive counselling in people with psychosis: study protocol for a randomised clinical trial. Trials 2022;23(1):773 [FREE Full text] [doi: 10.1186/s13063-022-06683-1] [Medline: 36100943]
72. Cloitre M, Petkova E, Wang J, Lu Lassell F. An examination of the influence of a sequential treatment on the course and impact of dissociation among women with PTSD related to childhood abuse. Depress Anxiety 2012;29(8):709-717. [doi: 10.1002/da.21920] [Medline: 22550033]
73. Paulik G, Newman-Taylor K, Steel C, Arntz A. Managing dissociation in imagery rescripting for voice hearers with trauma: lessons from a case series. Cognit Behav Pract 2022;29(2):434-445 [FREE Full text] [doi: 10.1016/j.cbpra.2020.06.009]
74. Davidson L, Bellamy C, Guy K, Miller R. Peer support among persons with severe mental illnesses: a review of evidence and experience. World Psychiatry 2012;11(2):123-128 [FREE Full text] [doi: 10.1016/j.wpsyc.2012.05.009] [Medline: 22654945]
75. Nelson C, Lusk R, Cawood C, Boore L, Ranganathan A, Lyubkin M. Predictors of CBT-pretreatment intervention engagement and completion: evidence for peer support. Psychol Serv 2019;16(3):381-387. [doi: 10.1037/ser0000268] [Medline: 30382747]
76. Ward T, Ball H, Montague A, Xanidis N, Myrie C, Murcutt I, et al. Relational work with distressing voices mirroring experiences of discrimination and marginalisation: three illustrative cases of an extended form of AVATAR therapy (AV-EXT). Psychol Psychother 2025. [doi: 10.1111/papt.70027] [Medline: 41355006]

Abbreviations

AV-BRF: AVATAR-Brief, a 6-session version of AVATAR therapy

AV-EXT: extended (12-session) form of AVATAR therapy
IPA: interpretative phenomenological analysis
NHS: National Health Service
PPI: patient and public involvement
RCT: randomized controlled trial
TA: therapeutic alliance
UCL: University College London

Edited by K Kahane; submitted 15.May.2025; peer-reviewed by H Griffiths, E Page, R Morais; comments to author 26.Aug.2025; revised version received 21.Nov.2025; accepted 02.Dec.2025; published 28.Jan.2026.

Please cite as:
Rutter-Eley E, Craig T, Garety P, Rus-Calafell M, Ball H, Clancy M, McDonnell J, Gumley A, Haddock G, Bucci S, Fornells-Ambrojo M, Baldwin N, Harling J, Phiri A, MacKenzie-Nash C, Hamilton N, Grant A, Edwards C, Ward T
"It Felt Good to Be Able to Say That Out Loud"—Therapeutic Alliance and Processes in AVATAR Therapy for People Who Hear Distressing Voices: Peer-Led Qualitative Study
JMIR Ment Health 2026;13:e77566
URL: https://mental.jmir.org/2026/1/e77566
doi:10.2196/77566
PMID:

Using Smartphone-Tracked Behavioral Markers to Recognize Depression and Anxiety Symptoms: Cross-Sectional Digital Phenotyping Study

George Aalbers[1,2], PhD; Andrea Costanzo[3], PhD; Raj Jagesar[3], PhD; Femke Lamers[2], PhD; Martien J H Kas[3], PhD; Brenda W J H Penninx[1,2], PhD

[1]Department of Psychiatry, Amsterdam University Medical Center, Vrije Universiteit, Oldenaller 1, Amsterdam, The Netherlands

[2]Mental Health Program, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands

[3]Groningen Institute for Evolutionary Life Sciences Faculty of Science and Engineering, University of Groningen, Groningen, The Netherlands

Corresponding Author:
George Aalbers, PhD
Department of Psychiatry, Amsterdam University Medical Center, Vrije Universiteit, Oldenaller 1, Amsterdam, The Netherlands

Abstract

Background: Depression and anxiety are prevalent but commonly missed and misdiagnosed, an important concern because many patients do not experience spontaneous recovery, and the duration of untreated illness is associated with worse outcomes.

Objective: This study aims to explore the potential of using smartphone-tracked behavioral markers to support diagnostics and improve recognition of these disorders.

Methods: We used the dedicated Behapp digital phenotyping platform to passively track location and app usage in 217 individuals, comprising symptomatic (n=109; depression/anxiety diagnosis or symptoms) and asymptomatic individuals (n=108; no diagnosis/symptoms). After quantifying 46 behavioral markers (eg, % time at home), we applied a machine learning approach to (1) determine which markers are relevant for depression/anxiety recognition and (2) develop and evaluate diagnostic prediction models for doing so.

Results: Our analysis identifies the total number of GPS-based trajectories as a potential marker of depression/anxiety, where individuals with fewer trajectories are more likely to be symptomatic. Models using this feature in combination with demographics or in isolation outperformed demographics-only models (area under the receiver operating characteristic curve$_{Mdn}$=0.60 vs 0.60 vs 0.51).

Conclusions: Collectively, these findings indicate that smartphone-tracked behavioral markers have limited discriminant ability in our study but potential to support future depression/anxiety diagnostics.

(JMIR Ment Health 2026;13:e80765) doi:10.2196/80765

KEYWORDS

mobile health; mobile phone; digital phenotype; digital biomarker; machine learning

Introduction

Depression and anxiety disorders commonly are not recognized by general practitioners (eg, for depression, sensitivity=47.3% - 50.1%; [1]). In practice, this diagnostic issue is an important concern because many patients do not experience (short-term) spontaneous recovery [2,3], and the duration of untreated illness is associated with worse outcomes [4]. Mounting evidence suggests depression/anxiety recognition might be improved by diagnostic prediction models that rely on smartphone-tracked behavioral markers such as homestay and app use (eg, for depression, sensitivity=72.5% - 75.0%) [5]. However, this research area—here referred to as digital phenotyping—is considered to be in its infancy [6]. More work is required to identify informative behavioral markers and evaluate their potential diagnostic utility for health care professionals.

Digital phenotyping refers to "moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices, in particular smartphones" [7]. By accurately and unobtrusively capturing mental illness dimensions in daily life (eg, sleep, social behavior), digital phenotyping could contribute to more precise disease stratification in the long run (ie, deep phenotyping [8]). However, an important first step is to evaluate if digital phenotyping can help us broadly distinguish individuals with and without depression/anxiety, above and beyond demographic features known to predict these symptoms (eg, age, sex, and years of education [9]). We here consider the potential use of smartphone-tracked location and app use for depression/anxiety recognition; other digital phenotyping data sources (eg, Bluetooth, accelerometer, light sensor) are beyond the scope of this article.

Theoretically, digital phenotyping should (to some extent) help distinguish symptomatic from asymptomatic individuals. By using the smartphone to continuously log individuals' GPS-based location and phone use, we can quantify smartphone-tracked behavioral markers such as time spent at home [10-13] that overlap or correlate with psychopathological symptoms [14]. For instance, specific anxiety disorders (eg, agoraphobia, social phobia) are defined by avoidance of specific contexts, and therefore, we might reasonably expect individuals with these symptoms to spend less time at leisure places and more time at home [15]. Similarly, smartphone app use might be relevant as it captures information about a person's social activity (eg, time spent on communication apps such as WhatsApp) [16] and sleep patterns [17,18], both of which are altered in depression/anxiety.

In the past decade, digital phenotyping research has provided evidence that passively logged location and smartphone use might be promising for depression/anxiety recognition. One relatively stable finding in the domain is that depression and anxiety are related to reduced locational variability (eg, lower variance and entropy, lower number of places visited, longer homestay) [14,19-22]. Research further suggests smartphone log data might contain diagnostically useful features [23]. For instance, some—although limited—evidence indicates depression might be indicated by greater duration [19,23] and entropy of smartphone use [24], and that increased social media and communication app use might predict momentary subjective stress [25]. Collectively, evidence indicates digital phenotyping data might have diagnostic utility.

An important limitation of digital phenotyping research remains that applications in clinical samples are relatively uncommon [26]. This is not surprising because these are more costly and difficult to investigate than convenience samples, but it is problematic because the envisioned use case of digital phenotyping is clinical [27]. The unique contribution of our study is that we analyze a sample of individuals with (n=109; symptomatic group) and without clinically relevant depression/anxiety symptoms (n=108; asymptomatic group) in whom up to 43 days of digital phenotyping data were collected with the Behapp platform [12,16,28,29] in the Netherlands Study of Depression and Anxiety (NESDA) [30].

Using an explainable artificial intelligence (XAI) approach, which has growing popularity in the domain (Shapley additive explanations [SHAP] [31]; eg, [19]), we aim to (1) identify which behavioral markers are indicative of depression/anxiety and explore the strength and nature of this relation and (2) develop and evaluate machine learning (ML) models that use these markers to recognize depression/anxiety. Notably, as our sample size is limited, we develop and evaluate this model not for model deployment in clinical practice but rather as an exploration to inform future, larger studies. Where applicable, we report in line with the recently published TRIPOD-AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis Plus Artificial Intelligence) guidelines [32].

Methods

Data

We here analyze data collected in the NESDA [30], as part of the Stress in Action consortium project [33]. This study uses the method of Penninx et al [30], and the method description partly reproduces their wording. NESDA participants were initially included for a baseline assessment with clinical interviews and surveys (2004 - 2007) and assessed for the seventh time at the 15-year follow-up (2019 - 2023). NESDA was designed to be representative of individuals with depressive and anxiety disorders in different health care settings and stages of the developmental history. Initially, participants were recruited from mental health care organizations, primary care, and the community setting. Participants were eligible if they were between 18 and 65 years; fluent in Dutch; and did not meet criteria for psychotic disorder, obsessive-compulsive disorder, bipolar disorder, or severe addiction disorder.

Specially trained clinical research staff conducted the composite international diagnostic interview [34] to determine if participants met *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (*DSM-IV*) criteria for depression and anxiety disorders, and participants completed a battery of self-report surveys using depression and anxiety symptom measures, including the Inventory of Depressive Symptomatology (IDS) [35] and Beck Anxiety Inventory [36]. A subset of NESDA participants installed a digital phenotyping app (Behapp, for more information see [11,28,29]) on their smartphone and provided the app permissions to continuously log their location (longitude and latitude) and smartphone app use (timestamps of when a specific app was opened and closed). On average, Behapp was activated on the day of the interview (SD 4 d), and self-report surveys were completed an average of 12 days before app activation.

Participants

We enrolled a total of 405 participants in the NESDA digital phenotyping study, 343 of whom had both clinical and digital phenotyping data (n=62 without any digital phenotyping data, who were excluded). Because iOS disallows app logging and this is an essential data source for our study, we excluded individuals with this operating system (n=24). Further, to ensure digital phenotyping data quality, we excluded individuals with fewer than 7 days of both app and GPS-based location data (n=102). Hence, all analyses were conducted in 217 participants. For a more extensive description of missingness patterns and a demographic comparison between iOS and Android users, see Multimedia Appendix 1. The overall sample size was determined by feasibility constraints (ie, we included as many participants as possible and we retained those with sufficient available data) rather than sample size calculation. We applied the commonly used 80/20 train-test split to determine the sample size for model development (training) and evaluation (testing). Power analysis using powerROC [37] showed that our test set sample size (217 * 0.20 arriving at 43-44 participants) is sufficient to confirm an area under the receiver operating characteristic curve

(AUROC)≥0.80 (prevalence of events in the test set=0.5, target width for estimated AUROC 95% CI <0.60).

Ethical Considerations

The NESDA study, including its digital phenotyping substudy, was approved by the Amsterdam UMC medical ethical committee (reference number 2003‐183). All participants provided informed consent for both clinical assessment and digital phenotyping.

For participation in a face-to-face assessment wave, respondents received a €15 (US $17.53) gift certificate and reimbursement for travel expenses in appreciation of their time and cooperation. All data were collected and processed in compliance with the General Data Protection Regulation (GDPR). To ensure participant privacy, we present only statistical aggregates that do not contain any personally identifiable information.

Outcome

DSM-IV–based diagnoses of depressive disorders (dysthymia and major depressive disorder [MDD]) and anxiety (social anxiety disorder, panic disorder with and without agoraphobia, agoraphobia, and generalized anxiety disorder) were established with the Composite International Diagnostic Interview (version 2.1 [34]), either in person or via phone call. Depression and anxiety symptom severity were assessed with the 30-item IDS [35] and BAI [36]. Outcome assessment was consistent across demographic groups. To evaluate if digital phenotyping data can broadly differentiate between individuals with and without symptoms, we combined these measures to form a binary outcome variable: symptomatic and asymptomatic. Symptomatic individuals had at least 1 depressive or anxiety disorder diagnosis in the past 6 months or an IDS or a BAI score exceeding thresholds specified in the survey manuals (IDS>13, BAI>9). Asymptomatic individuals did not have a diagnosis, and both IDS and BAI scores were below this threshold.

Predictors

We used the digital phenotyping platform Behapp to passively collect smartphone-based data without storing any content of web queries, messages, or calls, in compliance with the GDPR [38]. The Behapp app has already been successfully used to investigate neuropsychiatric phenotypes [10,11,13,16] and to measure behavioral changes during the COVID-19 pandemic [12]. In this study, the collected raw data consisted of GPS-based location and foreground app usage data. We sampled the participants’ latitude and longitude at least every 10 minutes (with higher sampling frequencies during movement). For foreground app usage, we logged when an individual opened and closed a specific app. The Behapp itself is an app that runs in the background and is only accessed for setting up data collection and to restart the app when no data are being collected.

Using the Behapp feature extraction pipeline, these raw data were used to compute features (ie, measurable quantities), such as total phone usage in hours per day or the percentage of time spent at home. Table 1 provides a synthetic overview of these features. Prior to model building, we applied data-driven feature selection (see below) to all available features and trained models using only the selected features. Note that although the feature ‘app addiction’ captures information that we believe conceptually maps onto app addiction, it is unclear how it relates to validated addiction surveys or addiction diagnoses.

Table . Synthetic overview of Behapp digital phenotyping features.

Feature group and features	Definition	Example
Location		
Number of (unique) stay points	Number of (unique) stay points (average per day). A stay point is a location where participants stay within a range of 150 m for more than 30 min. In the count of the unique stay points, repeated visits of the same stay point (eg, office) count as 1 visit or can be smaller than 1 when individual has a single stay point for most of data collection.	• Day 1: Home, work, gym • Day 2: Home, work, café • Day 3: Home, theater, park • ... • Number of stay points (total)=9 • Number of stay points (average per day)=9/3=3 • Number of unique stay points (total)=6 • Number of unique stay points (average per day)=6/3=2
Time spent stationary	Time spent at stay points (in min).	• Day 1: Home (10 h), work (8 h), gym (1 h) • Day 2: Home (13 h), work (9 h), café (2 h) • Time spent at stay points (average per day)=(10 + 8 + 1) + (13 + 9 + 2)/2 = 21.5
Home	Most frequently visited stay point of the top 3 stay points where most time was spent at night.	• Apartment a: 16 h • Apartment b: 9 h, 12 h, 10 h, 11 h • Night club location: 7 h • Apartment b=>Home
Trajectories	Each set of location data points in between stay points is saved separately as a trajectory if they contain a minimum of 20 data points (totaling at least 30 min).	• Location 1 (2 h), travel (35 min), location 2 (3 h)=>trajectory identified • Location 1 (2 h), travel (25 min), location 2 (3 h)=>trajectory not identified
App use		
App frequency	Number of times apps (in categories) were opened per day.	• WhatsApp is categorized as a communication app and Instagram as a social media app • Day 1: WhatsApp from 10:00 to 10:01, WhatsApp from 11:00 to 11:01, and Instagram from 11:30 to 11:40 • Day 2: WhatsApp from 10:00 to 10:01 • Number of times communication apps opened=3/2=1.5 • Number of times social media apps opened=1/2=0.5
App frequency at night	Number of times apps (in categories) were opened at night (between 00:00 and 05:00).	• —[a]
App duration	Duration (sum and mean) for which apps (in categories) were opened per day.	• Duration (sum, average per day) communication apps = (2 + 1)/2 = 1.5 min • Duration (mean, average per day) communication apps = (1 + 1)/2 = 1 min

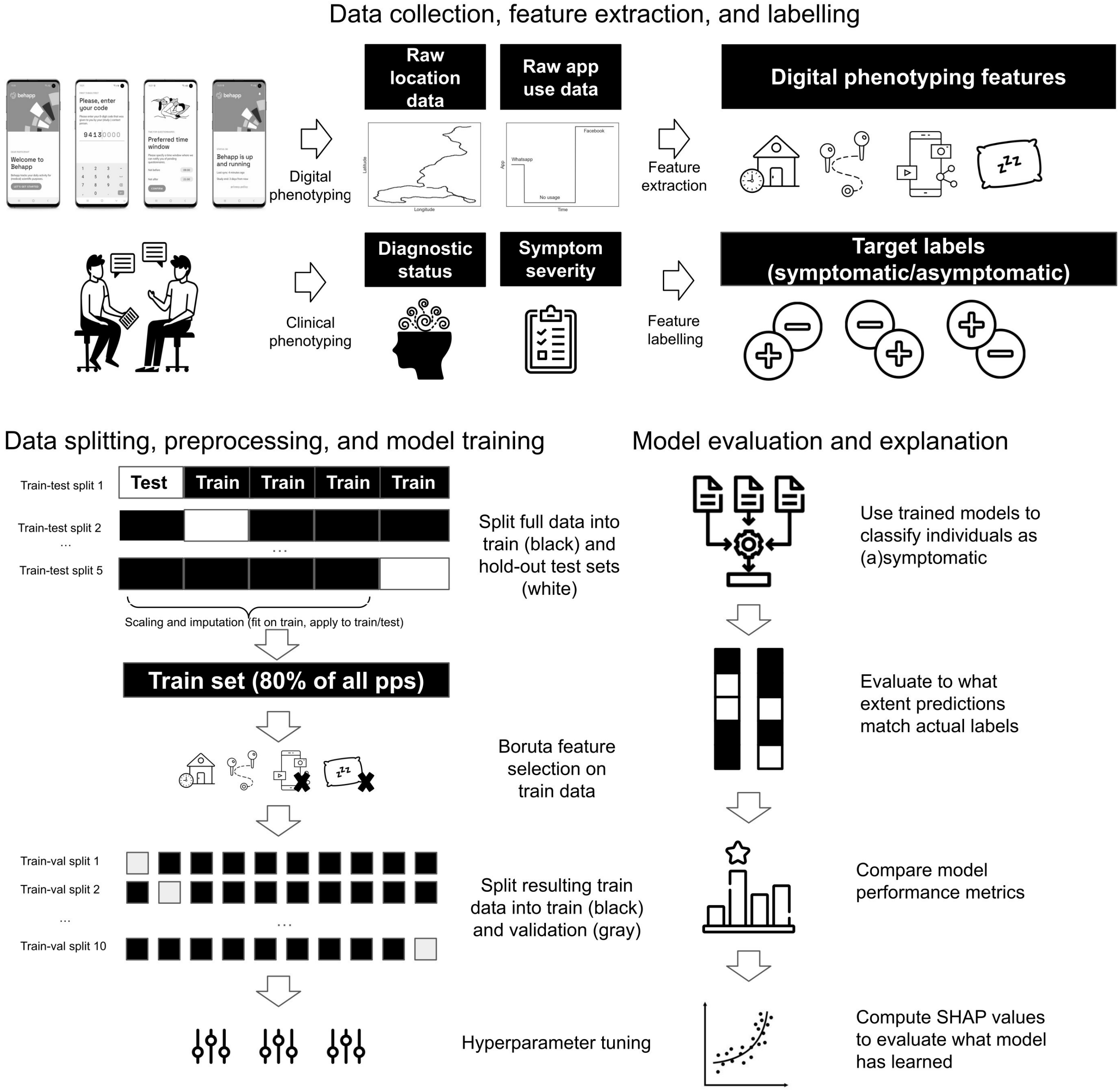
Feature group and features	Definition	Example
App duration at night	Duration (sum and mean) for which apps (in categories) were opened at night (between 00:00 and 05:00).	• —
App addiction	A value between 0 and 1 where 1 means that in each time interval of 20 min apps have been used at least once (average per day).	• WhatsApp from 10:00 to 10:01 • WhatsApp from 11:00 to 11:01 • WhatsApp from 12:00 to 12:01 WhatsApp from 13:00 to 13:01 • No other usage until 14:00 • Addiction = 4/(4 * 3) = 0.33 • 4*3 represents the number of 20-min intervals between 10:00 and 14:00

[a]Not applicable.

Analytical Methods

Figure 1 visualizes our ML pipeline (see [39] for all required Python code). After extracting features from the raw location and app usage data, we randomly split features and their corresponding psychiatric labels into 5 partitions, each containing (n=43) data points, corresponding to 20% of our total sample. To uniformly distribute individuals with and without depression/anxiety symptoms across data partitions, we apply a stratified 5-fold data split. We then iteratively select 4 partitions (referred to as the training set) and use these for minimum-maximum feature scaling, missing feature value imputation, feature selection, and hyperparameter tuning (with 10-fold stratified cross-validation) of a linear (ElasticNet logistic regression [LR]) and tree-based ML model (random forest [RF]) on each of 3 feature (sub)sets (ie, demographic, digital phenotyping, combined), after which we evaluate models on the remaining data partition (referred to as the test set). To impute missing feature values, we computed the mean of each feature in a training set and replaced the missing values in both this training set and its associated test set with this value. This procedure was repeated for all train-test set pairs. We repeat all steps until each data partition has been held out of model training once. Finally, for all trained models, we compute and visualize SHAP [31] values to clarify how models make their predictions based on specific feature values.

Figure 1. Visualization of the study design and machine learning pipeline. SHAP: Shapley additive explanations.



Feature Selection

To reduce our initial feature set to a subset of potentially informative items, we applied the well-known Boruta feature selection algorithm [40], which selects an all-relevant subset of features by first reshuffling the original features into so-called “shadow features” and training RF classifiers to determine if the former are more informative about the target than the latter. The added value of RF, a tree-based ensemble model [41], over a traditional linear model is the capacity to learn discontinuous, interactive associations without making assumptions (eg, absence of multicollinearity) that are likely to be violated in digital phenotyping data (eg, multicollinearity due to feature similarity).

Model Training

Using the Python library Scikit-learn [42] (version 1.5), we applied grid search stratified 10-fold cross-validation to tune hyperparameters of elasticNet-regularized LR [43] and RF [41] to maximize the model’s AUROC in validation data. We maximize the AUROC as this metric is typically used to assess how well a diagnostic prediction model can differentiate between individuals with and without a certain health outcome [32].

Model Evaluation

To estimate how well models might differentiate between individuals with and without depression/anxiety symptoms in a real-world setting, we let trained models make predictions on the hold-out test data (ie, 20% patients) and then evaluate to what extent they can correctly classify individuals with and without depression/anxiety. We used Scikit-learn (version 1.5) to compute evaluation metrics applied in related work [19] (accuracy, AUROC, F1, precision, and recall, computing F1, precision, and recall separately for asymptomatic (F10, Precision0, Recall0) and symptomatic individuals (F11, Precision1, Recall1). Because we use 5-fold nested cross-validation—which means we train models on five train-test splits—we also evaluate each trained model on 5 hold-out test sets. For each trained model, we provide the median score for each evaluation metric. To determine which model performed best, we use the AUROC as our primary evaluation metric, as this is typically done for binary classification tasks [32].

For theory-driven researchers and clinicians, an important limitation of the RF classifier is that this model does not have the interpretable parameters that make up linear models. We therefore explain our models using the Python library SHAP (version 0.46.0) to compute and visualize SHAP [31] values as a beeswarm plot. A beeswarm plot visualizes how changes in feature values affect probabilities output by the model. This visualization might be thought of as a visual stand-in for parameter estimates in linear models.

Results

Descriptives

Demographics were similar in the 2 groups, although female participants were overrepresented in the symptomatic group (Table 2). Relatively few individuals in the symptomatic group had a current diagnosis, with MDD being the most common diagnosis (n=26), followed by social phobia (n=16). However, by design, self-reported depression and anxiety symptoms were higher in the symptomatic than in the asymptomatic group. The median individual had 42 days of GPS-based location data and up to 43 days of app usage data. On average, symptomatic and asymptomatic individuals differed most strongly in their total number of leisure stay points, number of trajectories, duration of entertainment apps, and number of apps used (Multimedia Appendix 2). Most feature distributions were nonnormal, with the highest densities generally at the left tail. Feature distributions strongly overlap between symptomatic and asymptomatic individuals. Minor distributional differences are visible for features that quantify locational variability (ie, total time spent stationary, percentage of stay points visited once, total number of trajectories, total number of stay points, mean time spent stationary, total time traveled), and communication app use (Figure 2).

Table . Demographic and clinical descriptives for the asymptomatic and symptomatic groups.

Domain and Variable	Asymptomatic (n=108)	Symptomatic (n=109)
Demographics		
Age (years), mean (SD)	55.08 (12.82)	53.39 (12.24)
Years of education, mean (SD)	13.86 (2.88)	13.27 (3.28)
Sex (female), n (%)	65 (60.19)	78 (70.64)
Diagnosis in past 6 mo, n (%)		
Agoraphobia	—[a]	5 (4.59)
Panic disorder	—	12 (11.01)
Generalized anxiety disorder	—	6 (5.50)
Social phobia	—	16 (14.68)
Dysthymia	—	6 (5.50)
MDD[b]	—	26 (23.85)
Symptom severity, mean (SD)		
IDS[c] total	6.22 (3.59)	21.63 (8.40)
BAI[d] total	2.84 (2.50)	11.00 (6.35)
Data availability (days), median (IQR; range; SD)		
Location	42.00 (25.75-43; 7-43; 11.53)	42.00 (30-43; 7-43; 10.09)
App usage	42.50 (25-43; 8-43; 11.77)	43.00 (32-43; 8-43; 10.09)

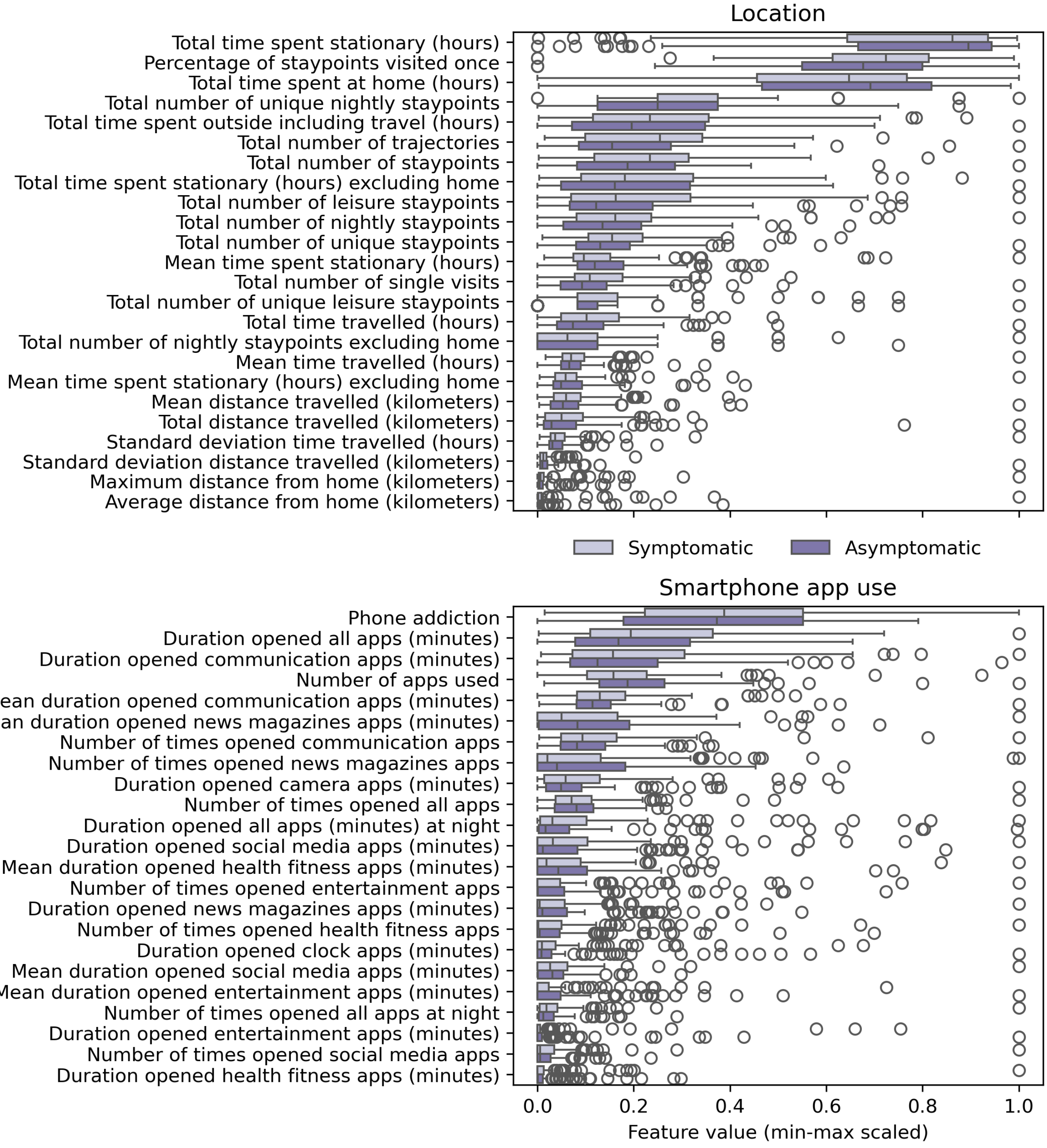
[a]Not applicable.

[b]MDD: major depressive disorder.

[c]IDS: Inventory of Depressive Symptomatology.

[d]BAI: Beck Anxiety Inventory.

Figure 2. Multiboxplot representing distributions of minimum and maximum scaled digital phenotyping features for each group (symptomatic=dark purple and asymptomatic=light purple). Plots are categorized by feature group (upper panel location, lower panel smartphone app use) and sorted by overall feature mean.



Feature Selection

The total number of GPS-based trajectories was selected in all train-test splits. Other features, the majority of which were locational (n=7 vs n=3 app use), were generally selected in only 1 data split (eg, mean duration of communication app use). Therefore, with respect to the full dataset and relative to all the other features, a person's number of location trajectories appears to most reliably indicate depression/anxiety. After an initial round of training and evaluating models, we observed suboptimal model performance, which we attributed to overfitting on feature selection. We therefore decided to retrain all models with the total number of GPS-based trajectories as the only digital phenotyping feature, as this was clearly the most stable feature of the data, and we successfully improved model performance by doing so. We present model performance and explanations of these simplified models here. Please note that model performance could be inflated due to decisions informed by the hold-out test data.

Model Performance

To evaluate if features are predictive not only in the train but also in the hold-out test data, we trained 2 common ML model types (LR, RF) to recognize depression/anxiety from digital phenotyping and demographic data. We considered the predictive performance of models trained with only the total number of GPS-based trajectories as their feature and models using a combination of the total number of GPS-based trajectories and demographic features (age, sex, and years of education). We compared how well these models performed relative to two baseline models: a dummy model that uniformly outputs asymptomatic or symptomatic groups and models trained using only demographic features.

Considering median AUROC values as the primary metric, LR trained on only the total number of GPS-based trajectories performed best across train-test splits. AUROCs for this model ($AUROC_{Mdn}$=0.61, Table 3) exceeded those of both baseline models (dummy model $AUROC_{Mdn}$=0.50; LR demographics model $AUROC_{Mdn}$=0.52). LR trained on the combined feature groups ($AUROC_{Mdn}$=0.56) outperformed both baseline models but performed worse than the LR using GPS-based trajectories. Performance of RF models trained on combined feature groups ($AUROC_{Mdn}$=0.60) was equal to that of RF models trained on GPS-based trajectories ($AUROC_{Mdn}$=0.60) and better than both baselines (dummy model $AUROC_{Mdn}$=0.50; RF demographics model $AUROC_{Mdn}$=0.51).

Table . Model performance metrics (median-aggregated across hold-out test sets, range in parentheses).

Feature group and model	AUROC[a]	F10	F11	Precision0	Precision1	Recall0	Recall1	Accuracy
Baseline model, median (IQR; range)								
DM[b]	0.50 (0.50-0.50; 0.50-0.50)	0.47 (0.45-0.55; 0.45-0.56)	0.47 (0.45-0.52; 0.45-0.56)	0.45 (0.45-0.55; 0.45-0.55)	0.48 (0.45-0.52; 0.45-0.57)	0.48 (0.45-0.55; 0.45-0.57)	0.45 (0.45-0.52; 0.45-0.55)	0.47 (0.45-0.53; 0.45-0.56)
All, median (IQR; range)								
LR[c]	0.56 (0.51-0.65; 0.49-0.67)	0.49 (0.47-0.55; 0.46-0.63)	0.61 (0.57-0.61; 0.42-0.72)	0.58 (0.53-0.60; 0.45-0.75)	0.54 (0.52-0.58; 0.48-0.64)	0.48 (0.41-0.52; 0.41-0.55)	0.64 (0.64-0.71; 0.45-0.82)	0.56 (0.52-0.58; 0.47-0.68)
RF[d]	0.60 (0.60-0.61; 0.58-0.64)	0.60 (0.55-0.62; 0.46-0.62)	0.54 (0.53-0.60; 0.46-0.63)	0.56 (0.56-0.61; 0.50-0.62)	0.58 (0.53-0.62; 0.53-0.62)	0.64 (0.62-0.64; 0.36-0.71)	0.50 (0.45-0.59; 0.41-0.76)	0.57 (0.56-0.58; 0.51-0.61)
Digital phenotyping, median (IQR; range)								
LR	0.61 (0.60-0.61; 0.56-0.62)	0.57 (0.56-0.62; 0.5-0.62)	0.59 (0.57-0.60; 0.46-0.64)	0.58 (0.53-0.61; 0.52-0.65)	0.58 (0.54-0.62; 0.53-0.63)	0.64 (0.50-0.64; 0.48-0.67)	0.59 (0.55-0.59; 0.41-0.71)	0.61 (0.53-0.61; 0.52-0.61)
RF	0.60 (0.58-0.62; 0.52-0.62)	0.59 (0.58-0.61; 0.43-0.62)	0.47 (0.45-0.53; 0.43-0.57)	0.52 (0.52-0.58; 0.43-0.58)	0.56 (0.53-0.59; 0.45-0.60)	0.67 (0.64-0.68; 0.43-0.68)	0.45 (0.41-0.48; 0.36-0.55)	0.54 (0.52-0.58; 0.44-0.59)
Demographics, median (IQR; range)								
LR	0.52 (0.51-0.62; 0.48-0.63)	0.49 (0.46-0.51; 0.39-0.53)	0.60 (0.57-0.61; 0.56-0.70)	0.56 (0.53-0.60; 0.47-0.75)	0.54 (0.52-0.56; 0.50-0.59)	0.41 (0.41-0.41; 0.33-0.48)	0.64 (0.64-0.71; 0.64-0.86)	0.56 (0.52-0.56; 0.48-0.64)
RF	0.51 (0.51-0.55; 0.47-0.61)	0.50 (0.46-0.51; 0.40-0.52)	0.55 (0.51-0.57; 0.50-0.57)	0.53 (0.50-0.53; 0.44-0.55)	0.52 (0.50-0.54; 0.46-0.54)	0.48 (0.41-0.50; 0.36-0.52)	0.59 (0.55-0.59; 0.50-0.62)	0.51 (0.51-0.53; 0.45-0.55)

[a]AUROC: area under the receiver operating characteristic curve.

[b] DM: dummy model. Accuracy is balanced for any minor class imbalance.

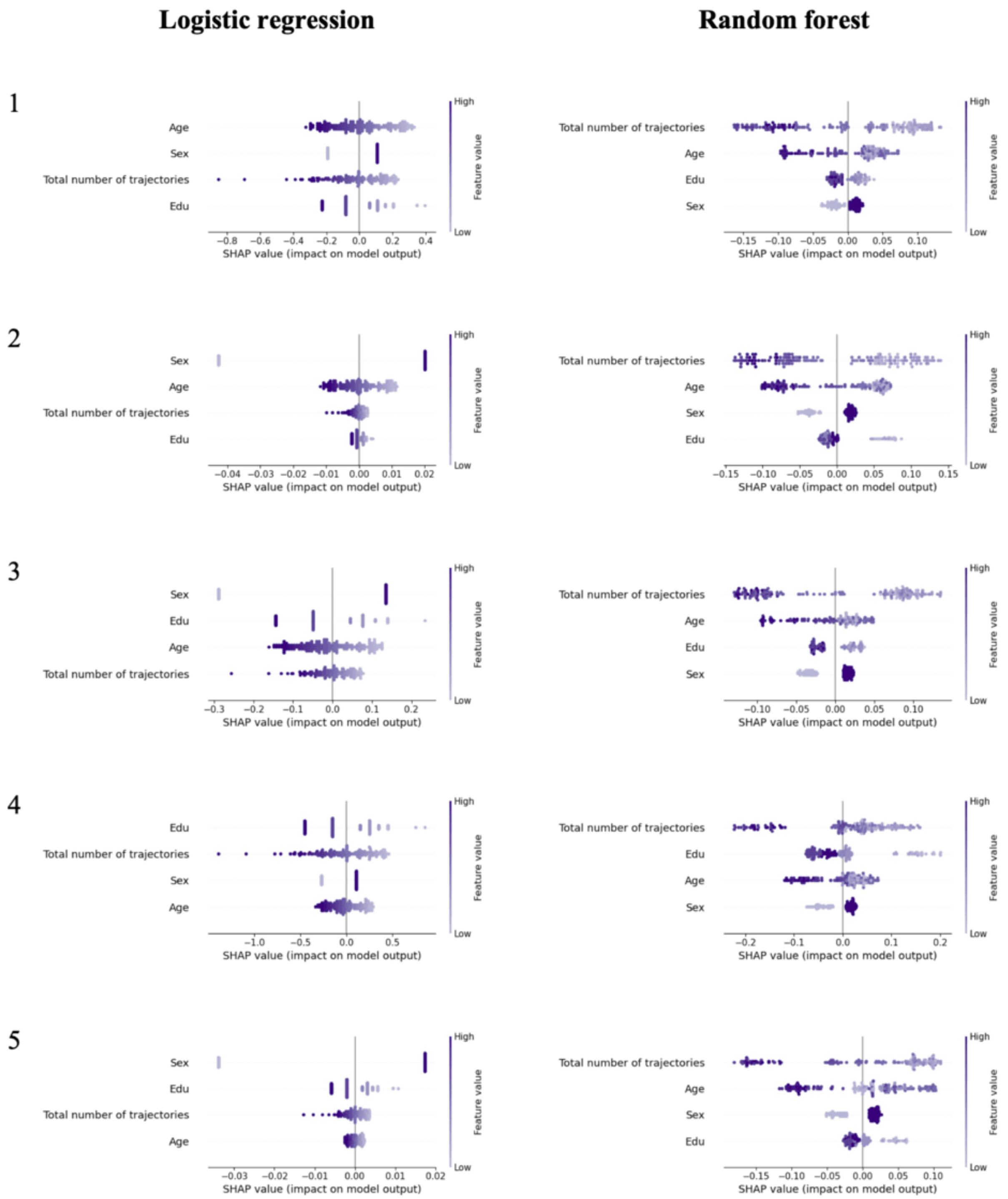
[c]LR: logistic regression.

[d]RF: random forest.

Model Explanation

As nonlinear ML models such as RF do not have directly interpretable parameters, we computed and visualized SHAP values as beeswarm plots. By inspecting these plots, we learn how ML models transform feature values into probabilities, and in doing so, we gain insight into the mapping from features to the outcome that the models have learned. Figure 3 shows that LR and RF models consistently learned a negative relation between an individual's total number of GPS-based trajectories per day and depression/anxiety. This means that for individuals with fewer trajectories, all models output a greater probability that they have depression/anxiety. Though capturing a relation with the same sign for this feature, LR and RF disagreed on its feature importance relative to demographic features. RF classifiers always assigned the highest feature importance to the total number of GPS-based locational trajectories, whereas LR always prioritized one or more demographic features.

Figure 3. Beeswarm plots for the two model types trained using sex, age, years of education, and GPS-based number of trajectories as features. SHAP: Shapley additive explanations.



Discussion

Principal Findings

To explore if digital phenotyping has the potential to support a diagnosis of depression/anxiety, we applied XAI to a unique combination of location, app use, and clinical assessment data collected in a subsample of NESDA participants with (n=109; symptomatic) and without depression/anxiety disorders or clinically relevant symptoms (n=108; asymptomatic). Our general findings suggest behavioral markers extracted from location and app use data potentially carry information about depression/anxiety, although their capacity to distinguish between those with and without clinically relevant symptoms is limited in the currently used data set.

Using a data-driven approach, we identify a number of GPS-based trajectories as a candidate behavioral marker for future studies. The total number of GPS-based trajectories, a locational feature that measures how frequently an individual moves between different stay points, was consistently selected across data splits, while other features were generally selected only once. Descriptive statistics and XAI analysis showed that individuals with fewer trajectories are more likely to be symptomatic and that this relation holds above and beyond demographic factors.

Our finding that individuals with fewer GPS-based trajectories are more likely to have depression/anxiety symptoms fits with previous empirical findings, providing more evidence that GPS-based behavioral features map onto depression/anxiety symptoms. A consistent finding in digital phenotyping has been that those with reduced locational variability (eg, lower variance and entropy, longer homestay) tend to have more depression and/or anxiety symptoms [19-22]. Conceptually, this behavioral feature fits with depression and anxiety symptoms that might diminish an individual's tendency to approach rewarding experiences (eg, anhedonia) or might reinforce their tendency to avoid negative experiences (eg, specific locations or situations such as social situations). Such symptoms potentially might cause individuals to get stuck in places (or rather to prevent them from getting unstuck), which would manifest itself as reduced GPS-based trajectories. Changes in GPS-based trajectories might have clinical use in terms of monitoring symptoms, but could also point to an intervention opportunity where individuals are encouraged to increase their daily number of trajectories.

Comparison With Prior Work

Model performance was less optimistic than in related ML work [19,23,24] but is consistent with statistically oriented studies that show weak relations between locational features and depression/anxiety [21,22]. Adequate study-to-study comparisons remain difficult to make, however, because digital phenotyping and clinical measures, sample characteristics, and modeling decisions differ from study to study and are likely to explain performance gaps. Hence, as suggested by others [44], an important avenue for digital phenotyping will be to harmonize study designs (eg, data collection, feature extraction, model types, cross-validation) to facilitate comparisons that are required to more adequately monitor progress in the domain. This call has been answered by academic consortia such as Stress in Action [33] that aim to collect digital phenotyping data at scale.

Model explanations using SHAP [31] showed our models have learned associations that are partially consistent with previous digital phenotyping studies. We found the total number of trajectories—which is conceptually similar to standard measures such as location variance or entropy—to be most important relative to all other Behapp features and negatively related to depression/anxiety, which matches previous findings [20,22,23,45] and well-known patterns that characterize depression/anxiety (eg, reduced motivation, social withdrawal [23]). However, contrary to previous work [24], our study did not identify any app use features as reliable predictors of depression/anxiety. Conceivably, this is because the relation between app use and depression/anxiety might hold in a specific subgroup only, as previous evidence suggests the association between app use and mental well-being potentially might differ from person to person [25,46-48].

Our findings are of interest not only for the development of diagnostic support systems but also for predict and preempt systems that aim to facilitate relapse prevention [49]. Diagnostic support systems aim to separate symptomatic from asymptomatic individuals based on differences between individuals (eg, symptomatic individuals tend to have fewer locational trajectories than asymptomatic individuals), while predict and preempt systems aim to identify onset of symptoms within an individual, based on behavioral differences between this individual's asymptomatic and symptomatic periods (eg, when an individual's trajectories start to decrease, they are increasingly symptomatic). Because our modeling approach, strictly speaking, is limited to between-subject conclusions, these findings do not necessarily imply that changes in the number of locational trajectories are indicative of symptom change. However, previous studies have already found within-person associations between locational features and depression/anxiety symptoms [21,22,50], indicating our findings might generalize from between-person to within-person and could potentially inform systems for relapse prevention.

Limitations

An important contribution of our work is that it investigated the potential use of digital phenotyping in a sample that included individuals with a current disorder. This is still relatively uncommon in digital phenotyping [26] as clinical samples are more difficult to study than convenience samples. Notwithstanding, our findings should be considered in light of the following limitations. Though larger than the average study in the domain (N=217 vs N_{Mean}=82) [51], our sample size was limited and had restricted demographics, in particular regarding age range. Combined with the fact that we excluded iOS users from our analysis, generalizability is limited to middle-aged Android users. However, considering the small demographic differences between Android and iOS users in our sample, this issue seems limited. Android users, on average, were somewhat older than iOS users, but did not differ in years of education and gender. Recent work in larger samples, however, has shown that Android ownership predicts lower levels of education, income, and extraversion [52], while other evidence suggests Android users are more likely to be men and older [53]. Future digital phenotyping work is needed for both iOS and Android users, as digital phenotyping screening tools ideally would be deployed irrespective of an individual's operating system (OS). Because technical architecture and privacy frameworks for a given OS might prevent certain data sources from being collected (eg, app use in iOS), this means digital phenotyping screening tools might need to be developed for each OS separately.

It is also important to note that—within the feasibility constraints on sample size that are very common in a clinical setting—we were unable to evaluate how digital phenotyping features relate to specific depression and anxiety disorder diagnoses. The sample we analyzed contained a limited number of individuals with current depression and anxiety diagnoses. However, because many individuals who experience substantial residual symptoms would be unsuitable to be included in a control group, we took this into account by combining participant diagnosis and symptom self-report. Further, we know from NESDA research reports that comorbidity between depression and anxiety disorders is high, especially when looking at lifetime prevalence. We have observed that about 60% to 80% of NESDA respondents have had depression and anxiety diagnoses [54] and, therefore, decided not to analyze them separately. However, symptom heterogeneity in our sample might have attenuated our ability to detect features that are relevant to specific disorders, such as social anxiety disorder. In a more homogeneous sample, other behavioral markers might be found to be relevant and, given that these behavioral markers could arguably map better onto specific symptom profiles, it is thinkable that model performance would be improved. We therefore encourage future work to consider comparisons of symptomatically homogeneous groups. Notwithstanding, research on heterogeneous samples such as our own is necessary to detect transdiagnostic smartphone-tracked behavioral markers.

Finally, we acknowledge that model performance might have been inflated as a result of data leakage. In an initial exploratory round of model evaluation on hold-out test data, we discovered model performance to be unstable across test sets (Multimedia Appendix 3). We attributed this to the potential overfitting feature selection in the individual training sets, which is a risk in small sample sizes. To stabilize model performance, we decided to only develop and evaluate models with the most consistently selected feature (ie, total number of GPS trajectories). Even though these generalized more reliably to hold-out test sets within our sample, it could be that the total number of GPS trajectories was consistently selected by chance and that our post hoc decision to only retain this feature might be tantamount to overfitting. Further, post hoc power calculations using powerROC [37] showed that, with the model performance in our study, our test sets were about three times smaller than what would be required to convincingly show models perform better than random guessing. It is therefore not surprising that follow-up DeLong model comparison tests showed no statistically significant differences in model performance nor that model evaluation metrics confidence intervals consistently overlapped with chance (Multimedia Appendix 4). All in all, these findings should be interpreted with caution and viewed as a first step that can inform larger-scale follow-up studies.

Future Research Directions

We recommend the following for future work that aims to develop a digital phenotyping-based symptom recognition system that can adequately differentiate between symptomatic and asymptomatic individuals. To ensure digital markers are consistently defined across studies, digital phenotyping studies would benefit from developing and adhering to an ontology of digital markers (for an example under development, see [55]). This ontology would ideally map digital markers (and configurations thereof) to specific symptoms or syndromes and could still include behavioral markers that were not marked as relevant in the present dataset, but have a strong conceptual mapping onto disorder definitions (eg, homestay is a clinical marker of agoraphobia). In a sufficiently large dataset with adequate diagnostic labels (conceivably in the order of thousands of participants), such an ontology could be used to develop multigroup classification models that can leverage digital markers to identify individuals as having no symptoms or (symptoms of) one or more specific disorders (eg, agoraphobia or agoraphobia with MDD).

It is highly recommended to design future digital phenotyping studies with model evaluation in mind, using a prior power analysis. Given the model performance in this study, post hoc sample size calculation suggests that at least 122 individuals should be held out of training for model evaluation, meaning that a sample size of over 600 individuals would have been needed for both training and evaluation. Of note, fewer individuals would be required for sufficient statistical power if model performance is improved substantially, which might possibly be achieved with greater symptom contrasts between groups (ie, comparing individuals without symptoms to individuals with severe symptoms) and greater symptom homogeneity within groups (ie, comparing individuals without symptoms to individuals with a specific disorder).

Development of a multimodal digital phenotyping toolkit, longitudinal measurement of much larger samples, and follow-up research on theoretically relevant markers is underway in the Stress in Action consortium [33]. We envision that, over time, this multimodal digital phenotyping toolkit might be used to trigger traditional symptom screening instruments such as the 9-item Patient Health Questionnaire (PHQ-9) when symptoms are most likely (for a similar approach with wearables, see [49]), given changes in an individual's digital phenotyping data. Symptom screening surveys have high sensitivity for detecting mental illness. For instance, the PHQ-9 has a sensitivity of 0.88 (95% CI 0.83 to 0.92) and a specificity of 0.85 (0.82-0.88) for MDD [56], which is unlikely to be outmatched by digital phenotyping models. In practice, however, individuals are unlikely to consistently complete symptom surveys for extended periods, which can be a significant burden. By using digital phenotyping, we might be able to help reduce this burden and potentially improve early symptom detection.

Conclusion

In all, digital phenotyping, here operationalized as passive logging of location and app use, offers insights into behavioral patterns that could potentially differentiate individuals with clinically relevant depression/anxiety symptoms from those without. In the unique NESDA sample comprising both symptomatic and asymptomatic individuals, we identified a specific smartphone-tracked behavioral marker, namely the total number of GPS-based trajectories, that may indicate these symptoms. Our findings align with previous studies suggesting ML models might be able to leverage smartphone-tracked behavioral markers to recognize symptomatic individuals. Although we show that such markers cannot support diagnostics on their own, we believe they are sufficiently promising to be considered in future deep phenotyping of depression and anxiety.

Acknowledgments

All the authors thank all Netherlands Study of Depression and Anxiety (NESDA) participants and research staff for their time and effort spent on this study.

Funding

GA, FL, and BWJHP are supported by Stress in Action. The research project Stress in Action [33] is financially supported by the Dutch Research Council and the Dutch Ministry of Education, Culture and Science (NWO gravitation grant number 024.005.010). AC, RJ, BWJHP and MJHK are supported by the PRISM 2 project. The PRISM 2 project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No. 101034377. The JU receives support from the European Union's Horizon 2020 research and innovation program, EFPIA, and Cohen Veterans Bioscience (CVB). The website reflects only the authors' views, and IMI JU, EFPIA, CVB, and the European Commission are not responsible for any use that may be made of the information it contains. The infrastructure for the NESDA study (www.nesda.nl) is funded through the Geestkracht program of the Netherlands Organization for Health Research and Development (ZonMw, grant number 10-000-1002) and financial contributions by participating universities and mental health care organizations (Amsterdam University Medical Centers, location VUmc; GGZ inGeest; Leiden University Medical Center; Leiden University; GGZ Rivierduinen; University Medical Center Groningen; University of Groningen; Lentis; GGZ Friesland; GGZ Drenthe; Rob Giel Onderzoekscentrum). The funding sources were not involved in study design, data collection, data analysis, data interpretation, writing of the report, or decision to submit the article for publication.

Data Availability

Research data cannot be shared.

Authors' Contributions

Conceptualization: GA, FL, BWJHP
Data curation: GA, AC, FL
Formal analysis: GA, AC
Funding acquisition: FL, MJHK, BWJHP
Investigation: FL
Methodology: GA, AC
Project administration: FL
Supervision: FL, BWJHP
Visualization: GA
Writing – original draft: GA, AC
Writing – review & editing: GA, AC, RJ, FL, MJHK, BWJHP

Conflicts of Interest

None declared.

Multimedia Appendix 1
Missingness patterns and operating system demographics.
[DOCX File, 30 KB - mental_v13i1e80765_app1.docx]

Multimedia Appendix 2
Descriptives for the top 5 digital phenotyping features.
[DOCX File, 26 KB - mental_v13i1e80765_app2.docx]

Multimedia Appendix 3
Model performance with initial exploratory Boruta feature selection.
[DOCX File, 967 KB - mental_v13i1e80765_app3.docx]

Multimedia Appendix 4
Statistical model comparisons (DeLong test) and bootstrapped CIs.
[XLSX File, 28 KB - mental_v13i1e80765_app4.xlsx]

References

1. Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. Lancet 2009 Aug 22;374(9690):609-619. [doi: 10.1016/S0140-6736(09)60879-5] [Medline: 19640579]
2. Mekonen T, Ford S, Chan GCK, Hides L, Connor JP, Leung J. What is the short-term remission rate for people with untreated depression? A systematic review and meta-analysis. J Affect Disord 2022 Jan 1;296:17-25. [doi: 10.1016/j.jad.2021.09.046] [Medline: 34583099]
3. Whiteford HA, Harris MG, McKeon G, et al. Estimating remission from untreated major depression: a systematic review and meta-analysis. Psychol Med 2013 Aug;43(8):1569-1585. [doi: 10.1017/S0033291712001717] [Medline: 22883473]
4. Ghio L, Gotelli S, Marcenaro M, Amore M, Natta W. Duration of untreated illness and outcomes in unipolar depression: a systematic review and meta-analysis. J Affect Disord 2014 Jan;152-154:45-51. [doi: 10.1016/j.jad.2013.10.002] [Medline: 24183486]
5. Moura I, Teles A, Viana D, Marques J, Coutinho L, Silva F. Digital phenotyping of mental health using multimodal sensing of multiple situations of interest: a systematic literature review. J Biomed Inform 2023 Feb;138:104278. [doi: 10.1016/j.jbi.2022.104278] [Medline: 36586498]
6. Roefs A, Fried EI, Kindt M, et al. A new science of mental disorders: Using personalised, transdiagnostic, dynamical systems to understand, model, diagnose and treat psychopathology. Behav Res Ther 2022 Jun;153:104096. [doi: 10.1016/j.brat.2022.104096] [Medline: 35500541]
7. Onnela JP. Opportunities and challenges in the collection and analysis of digital phenotyping data. Neuropsychopharmacology 2021 Jan;46(1):45-54. [doi: 10.1038/s41386-020-0771-3] [Medline: 32679583]
8. Robinson PN. Deep phenotyping for precision medicine. Hum Mutat 2012 May;33(5):777-780. [doi: 10.1002/humu.22080] [Medline: 22504886]
9. Licht CMM, de Geus EJC, Zitman FG, Hoogendijk WJG, van Dyck R, Penninx BWJH. Association between major depressive disorder and heart rate variability in the Netherlands Study of Depression and Anxiety (NESDA). Arch Gen Psychiatry 2008 Dec;65(12):1358-1367. [doi: 10.1001/archpsyc.65.12.1358] [Medline: 19047522]
10. Sverdlov O, Curcic J, Hannesdottir K, et al. A study of novel exploratory tools, digital technologies, and central nervous system biomarkers to characterize unipolar depression. Front Psychiatry 2021;12:640741. [doi: 10.3389/fpsyt.2021.640741] [Medline: 34025472]
11. Jongs N, Jagesar R, van Haren NEM, et al. A framework for assessing neuropsychiatric phenotypes by using smartphone-based location data. Transl Psychiatry 2020 Jul 1;10(1):211. [doi: 10.1038/s41398-020-00893-4] [Medline: 32612118]
12. Jagesar RR, Roozen MC, van der Heijden I, et al. Digital phenotyping and the COVID-19 pandemic: capturing behavioral change in patients with psychiatric disorders. Eur Neuropsychopharmacol 2021 Jan;42:115-120. [doi: 10.1016/j.euroneuro.2020.11.012] [Medline: 33298386]
13. Kas MJH, Jongs N, Mennes M, et al. Digital behavioural signatures reveal trans-diagnostic clusters of schizophrenia and Alzheimer’s disease patients. Eur Neuropsychopharmacol 2024 Jan;78:3-12. [doi: 10.1016/j.euroneuro.2023.09.010] [Medline: 37864982]
14. Wang R, Wang W, Dasilva A, et al. Tracking depression dynamics in college students using mobile phone and wearable sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Mar;2(1):43. [doi: 10.1145/3191775] [Medline: 39449996]
15. Boukhechba M, Chow P, Fua K, Teachman BA, Barnes LE. Predicting social anxiety from global positioning system traces of college students: feasibility study. JMIR Ment Health 2018 Jul 4;5(3):e10101. [doi: 10.2196/10101] [Medline: 29973337]
16. Muurling M, Reus LM, de Boer C, et al. Assessment of social behavior using a passive monitoring app in cognitively normal and cognitively impaired older adults: observational study. JMIR Aging 2022 May 20;5(2):e33856. [doi: 10.2196/33856] [Medline: 35594063]
17. Aledavood T, Kivimäki I, Lehmann S, Saramäki J. Quantifying daily rhythms with non-negative matrix factorization applied to mobile phone data. Sci Rep 2022 Apr 1;12(1):5544. [doi: 10.1038/s41598-022-09273-y] [Medline: 35365710]
18. Borger JN, Huber R, Ghosh A. Capturing sleep-wake cycles by using day-to-day smartphone touchscreen interactions. NPJ Digit Med 2019;2(1):73. [doi: 10.1038/s41746-019-0147-4] [Medline: 31372507]
19. Opoku Asare K, Moshe I, Terhorst Y, et al. Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status: A longitudinal data analysis. Pervasive Mob Comput 2022 Jul;83:101621. [doi: 10.1016/j.pmcj.2022.101621]
20. Moshe I, Terhorst Y, Opoku Asare K, et al. Predicting symptoms of depression and anxiety using smartphone and wearable data. Front Psychiatry 2021;12:625247. [doi: 10.3389/fpsyt.2021.625247] [Medline: 33584388]
21. Chow PI, Fua K, Huang Y, et al. Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students. J Med Internet Res 2017 Mar 3;19(3):e62. [doi: 10.2196/jmir.6820] [Medline: 28258049]
22. Zhang Y, Folarin AA, Sun S, et al. Longitudinal relationships between depressive symptom severity and phone-measured mobility: dynamic structural equation modeling study. JMIR Ment Health 2022 Mar 11;9(3):e34898. [doi: 10.2196/34898] [Medline: 35275087]

23. Saeb S, Zhang M, Karr CJ, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015 Jul 15;17(7):e175. [doi: 10.2196/jmir.4273] [Medline: 26180009]
24. Opoku Asare K, Terhorst Y, Vega J, Peltonen E, Lagerspetz E, Ferreira D. Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: exploratory study. JMIR Mhealth Uhealth 2021 Jul 12;9(7):e26540. [doi: 10.2196/26540] [Medline: 34255713]
25. Aalbers G, Hendrickson AT, Vanden Abeele MM, Keijsers L. Smartphone-tracked digital markers of momentary subjective stress in college students: idiographic machine learning analysis. JMIR Mhealth Uhealth 2023 Mar 23;11:e37469. [doi: 10.2196/37469] [Medline: 36951924]
26. De Angel V, Lewis S, White K, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ Digit Med 2022 Jan 11;5(1):3. [doi: 10.1038/s41746-021-00548-8] [Medline: 35017634]
27. Insel TR. Digital phenotyping: a global tool for psychiatry. World Psychiatry 2018 Oct;17(3):276-277. [doi: 10.1002/wps.20550] [Medline: 30192103]
28. Behapp. URL: www.behapp.com [accessed 2025-12-19]
29. Jagesar RR, Vorstman JA, Kas MJ. Requirements and operational guidelines for secure and sustainable digital phenotyping: design and development study. J Med Internet Res 2021 Apr 7;23(4):e20996. [doi: 10.2196/20996] [Medline: 33825695]
30. Penninx BWJH, Beekman ATF, Smit JH, et al. The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. Int J Methods Psychiatr Res 2008;17(3):121-140. [doi: 10.1002/mpr.256] [Medline: 18763692]
31. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Presented at: Proceedings of the 31st International Conference on Neural Information Processing Systems; Dec 4, 2017; Long Beach, CA p. 4768-4777. [doi: 10.5555/3295222.3295230]
32. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 16;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
33. Stress in Action. URL: http://www.stress-in-action.nl/ [accessed 2025-12-19]
34. Wittchen HU. Reliability and validity studies of the WHO--Composite International Diagnostic Interview (CIDI): a critical review. J Psychiatr Res 1994;28(1):57-84. [doi: 10.1016/0022-3956(94)90036-1] [Medline: 8064641]
35. Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH. The Inventory of Depressive Symptomatology (IDS): psychometric properties. Psychol Med 1996 May;26(3):477-486. [doi: 10.1017/s0033291700035558] [Medline: 8733206]
36. Beck AT, Epstein N, Brown G, Steer RA. An inventory for measuring clinical anxiety: psychometric properties. J Consult Clin Psychol 1988 Dec;56(6):893-897. [doi: 10.1037//0022-006x.56.6.893] [Medline: 3204199]
37. Grolleau F, Tibshirani R, Chen JH. powerROC: an interactive WebwTool for sample size calculation in assessing models' discriminative abilities. AMIA Jt Summits Transl Sci Proc 2025;2025:196-204. [Medline: 40502274]
38. Mulder T, Jagesar RR, Klingenberg AM, P Mifsud Bonnici J, Kas MJ. New European privacy regulation: assessing the impact for digital medicine innovations. Eur Psychiatry 2018 Oct;54:57-58. [doi: 10.1016/j.eurpsy.2018.07.003] [Medline: 30121506]
39. George-aalbers/cross-sectional-digital-phenotyping-study. Github. URL: https://github.com/george-aalbers/cross-sectional-digital-phenotyping-study [accessed 2025-01-14]
40. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw 2010;36(11). [doi: 10.18637/jss.v036.i11]
41. Breiman L. Random forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
42. Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(null):2825-2830. [doi: 10.5555/1953048.2078195]
43. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Stat Soc Series B 2005 Apr 1;67(2):301-320. [doi: 10.1111/j.1467-9868.2005.00503.x]
44. Rohani DA, Faurholt-Jepsen M, Kessing LV, Bardram JE. Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: systematic review. JMIR Mhealth Uhealth 2018 Aug 13;6(8):e165. [doi: 10.2196/mhealth.9691] [Medline: 30104184]
45. Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ 2016;4:e2537. [doi: 10.7717/peerj.2537] [Medline: 28344895]
46. Aalbers G, vanden Abeele MMP, Hendrickson AT, de Marez L, Keijsers L. Caught in the moment: are there person-specific associations between momentary procrastination and passively measured smartphone use? Mobile Media & Communication 2022 Jan;10(1):115-135. [doi: 10.1177/2050157921993896]
47. Verbeij T, Pouwels JL, Beyens I, Valkenburg PM. Experience sampling self-reports of social media use have comparable predictive validity to digital trace measures. Sci Rep 2022 May 9;12(1):7611. [doi: 10.1038/s41598-022-11510-3] [Medline: 35534600]
48. Siebers T, Beyens I, Valkenburg PM. The effects of fragmented and sticky smartphone use on distraction and task delay. Mob Med Commun 2024 Jan;12(1):45-70. [doi: 10.1177/20501579231193941]
49. Vairavan S, Rashidisabet H, Li QS, et al. Personalized relapse prediction in patients with major depressive disorder using digital biomarkers. Sci Rep 2023 Oct 30;13(1):18596. [doi: 10.1038/s41598-023-44592-8] [Medline: 37903878]

50. Canzian L, Musolesi M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. Presented at: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing; Sep 7-11, 2015. [doi: 10.1145/2750858.2805845]
51. Melcher J, Hays R, Torous J. Digital phenotyping for mental health of college students: a clinical review. Evid Based Ment Health 2020 Nov;23(4):161-166. [doi: 10.1136/ebmental-2020-300180] [Medline: 32998937]
52. Schoedel R, Reiter T, Krämer MD, et al. Person-related selection bias in mobile sensing research: robust findings from two panel studies. PsyArXiv. Preprint posted online on Sep 11, 2025. [doi: 10.31234/osf.io/9w7hu_v1]
53. Shaw H, Ellis DA, Kendrick LR, Ziegler F, Wiseman R. Predicting smartphone operating system from personality and individual differences. Cyberpsychol Behav Soc Netw 2016 Dec;19(12):727-732. [doi: 10.1089/cyber.2016.0324] [Medline: 27849366]
54. Lamers F, van Oppen P, Comijs HC, et al. Comorbidity patterns of anxiety and depressive disorders in a large cohort study. J Clin Psychiatry 2011 Mar 15;72(3):341-348. [doi: 10.4088/JCP.10m06176blu]
55. Ontology for digital markers in mental health (ODIM-MH). GitHub. URL: https://github.com/MentalHealthMission/MHM-ontology/tree/main [accessed 2025-12-31]
56. Levis B, Benedetti A, Thombs BD. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ 2019 Apr 9;365:l1476. [doi: 10.1136/bmj.l1476] [Medline: 30967483]

Abbreviations

AUROC: area under the receiver operating characteristic curve
BAI: Beck Anxiety Inventory
***DSM-IV*:** *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*
GDPR: General Data Protection Regulation
IDS: Inventory of Depressive Symptomatology
LR: logistic regression
MDD: major depressive disorder
ML: machine learning
NESDA: Netherlands Study of Depression and Anxiety
PHQ-9: 9-item Patient Health Questionnaire
RF: random forest
SHAP: Shapley additive explanations
TRIPOD-AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis Plus Artificial Intelligence
XAI: explainable artificial intelligence

Edited by J Torous; submitted 16.Jul.2025; peer-reviewed by A Bunevicius, A Tai; revised version received 13.Oct.2025; accepted 31.Oct.2025; published 26.Jan.2026.

Please cite as:
Aalbers G, Costanzo A, Jagesar R, Lamers F, Kas MJH, Penninx BWJH
Using Smartphone-Tracked Behavioral Markers to Recognize Depression and Anxiety Symptoms: Cross-Sectional Digital Phenotyping Study
JMIR Ment Health 2026;13:e80765
URL: https://mental.jmir.org/2026/1/e80765
doi:10.2196/80765

Original Paper

Stakeholder Perspectives on Humanistic Implementation of Computer Perception in Health Care: Qualitative Study

Kristin M Kostick-Quenet[1], PhD; Meghan E Hurley[1], MA; Syed Ayaz[1], MA; John D Herrington[2,3], PhD; Casey J Zampella[2,3], PhD; Julia Parish-Morris[2,3], PhD; Birkan Tunç[2,3], PhD; Gabriel Lázaro-Muñoz[4], PhD, JD; Jennifer Blumenthal-Barby[1], PhD; Eric A Storch[5], PhD

[1]Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX, United States

[2]Department of Child and Adolescent Psychiatry and Behavioral Sciences, Center for Autism Research, Children's Hospital of Philadelphia, Philadelphia, PA, United States

[3]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[4]Department of Neurosurgery, Massachusetts General Hospital, Boston, MA, United States

[5]Menninger Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, Houston, TX, United States

Corresponding Author:
Kristin M Kostick-Quenet, PhD
Center for Medical Ethics and Health Policy
Baylor College of Medicine
One Baylor Plaza
Suite 310D
Houston, TX, 77030
United States
Phone: 1 346 691 8949
Email: kristin.kostick@bcm.edu

Abstract

Background: Computer perception (CP) technologies—including digital phenotyping, affective computing, and related passive sensing approaches—offer unprecedented opportunities to personalize health care, especially mental health care, yet they also provoke concerns about privacy, bias, and the erosion of empathic, relationship-centered practice. At present, it remains elusive what stakeholders who design, deploy, and experience these tools in real-world settings perceive as the risks and benefits of CP technologies.

Objective: This study aims to explore key stakeholder perspectives on the potential benefits, risks, and concerns associated with integrating CP technologies into patient care. A better understanding of these concerns is crucial for responding to and mitigating such concerns via design implementation strategies that augment, rather than compromise, patient-centered and humanistic care and associated outcomes.

Methods: We conducted in-depth, semistructured interviews with 102 stakeholders involved at key points in CP's development and implementation: adolescent patients (n=20) and their caregivers (n=20); frontline clinicians (n=20); technology developers (n=21); and ethics, legal, policy, or philosophy scholars (n=21). Interviews (~ 45 minutes each) explored perceived benefits, risks, and implementation challenges of CP in clinical care. Transcripts underwent thematic analysis by a multidisciplinary team; reliability was enhanced through double coding and consensus adjudication.

Results: Stakeholders raised concerns across 7 themes: (1) Data Privacy and Protection (88/102, 86.3%); (2) Trustworthiness and Integrity of CP Technologies (72/102, 70.6%); (3) direct and indirect Patient Harms (65/102, 63.7%); (4) Utility and Implementation Challenges (60/102, 58.8%); (5) Patient-Specific Relevance (24/102, 23.5%); (6) Regulation and Governance (17/102, 16.7%); and (7) Philosophical Critiques of reductionism (13/102, 12.7%). A cross-cutting insight was the primacy of context and subjective meaning in determining whether CP outputs are clinically valid and actionable. Participants warned that without attention to these factors, algorithms risk misclassification and dehumanization of care.

Conclusions: To operationalize humanistic safeguards, we propose "personalized road maps": co-designed plans that predetermine which metrics will be monitored, how and when feedback is shared, thresholds for clinical action, and procedures for reconciling discrepancies between algorithmic inferences and lived experience. Road maps embed patient education, dynamic consent, and tailored feedback, thereby aligning CP deployment with patient autonomy, therapeutic alliance, and ethical transparency. This multistakeholder study provides the first comprehensive, evidence-based account of relational, technical, and governance challenges raised by CP tools in clinical care. By translating these insights into personalized road maps, we offer a practical framework for developers, clinicians, and policy makers seeking to harness continuous behavioral data while preserving the humanistic core of care.

(JMIR Ment Health 2026;13:e79182) doi:10.2196/79182

KEYWORDS

computer perception; digital phenotyping; ethics; humanistic care; artificial intelligence; stakeholder engagement; context; consent; affective computing

Introduction

Computer Perception Tools in Mental Health Care

Computer perception (CP) tools, including digital phenotyping, affective computing, computational behavioral analysis, and other approaches that entail continuous and passive data collection using wearables and smartphone sensing, have been positioned as a remedy for longstanding diagnostic and informational gaps in mental health care. The term “computer perception” references the artificial intelligence (AI) subfield of computer “vision” but acknowledges a wider range of perceptive modalities beyond vision alone (eg, “hearing” through microphones, motion detection through accelerometers), referring not only to sensory acquisition but also to a system’s capacity to interpret, classify, and act upon such data—analogous to human perceptual processing that integrates recognition and interpretation. By leveraging sensors already embedded in everyday devices, these systems promise scalable, accessible surveillance of behaviors, as well as mood, cognition, and sociability, potentially addressing medicine’s chronic reliance on infrequent patient self-reports and clinician observation to gain insights into psychosocial, behavioral, and physiological states [1,2]. Although this study centers on mental health care, the ethical and translational issues examined here (ie, around inference, interpretation, and the integration of perceptual data into care) extend to other domains of medicine where continuous data streams are increasingly used for diagnosis and decision-making. CP tools also promise a personalized and patient-tailored diagnostic and therapeutic approach, in line with precision medicine goals [3-5]. Early studies suggest that CP-derived markers can forecast relapse in bipolar disorder, detect prodromal psychosis, tailor just-in-time behavioral prompts, and potentially widen access to mental health care. Yet the very features that make CP appealing also expose patients to unprecedented privacy risks, algorithmic bias, and a potential erosion of empathic, relationship-centered care [1,6-8].

Ethicists, regulators, and frontline stakeholders caution that integrating such pervasive sensing into care can imperil core values of confidentiality, fairness, and relational trust [9-12]. These impacts can be exacerbated by opaque algorithms, unclear pathways for secondary data reuse, and difficulties in obtaining meaningful informed consent in continuous monitoring scenarios. A limited number of studies [13-15] provide a foundation for understanding some of these concerns; however, no empirical research to date offers a comprehensive view of the wide-ranging perspectives held by diverse stakeholder groups regarding the benefits and risks of integrating CP into care. This study addresses this gap through an empirical exploration of diverse stakeholder perspectives, with special attention to impacts on humanistic, relationship-centered care.

The rationale for focusing on humanistic care is to underscore that good care, whether technological or manual, depends on recognizing the patient as a person with values, context, and dignity. Humanistic and humanized care frameworks [16-18] remind us that respectful dialogue, cultural sensitivity, and patient partnership are interwoven into the moral fabric of good practice [19]. Whether CP ultimately augments or erodes that fabric depends on how well designers, clinicians, and regulators anticipate the spectrum of ethical concerns voiced by those who will build, deploy, or live with these systems. This study, therefore, turns to those diverse stakeholders—developers, clinicians, patients, caregivers, and ethics, legal, and policy scholars—to ask how their concerns can guide the integration of CP in ways that preserve, rather than diminish, the humanization of care. While mental health provides a particularly vivid setting for exploring these questions, the concerns articulated by participants resonate across many areas of health care and health technology innovation.

Background

What makes CP technologies unique is that they increasingly involve algorithmic *inferences* about a person’s moment-to-moment mental or sociobehavioral state, or about predicted outcomes such as mood relapse, suicidality, or treatment response [2,12,20,21]. These inferences are enabled by the ingestion of vast amounts of behavioral, physiological, and environmental signals from (usually) ordinary connected devices such as smartphones and wearables. Less often, they may involve implantable systems that continuously monitor physiological [22,23] or neural activity [24]. In psychiatric contexts, the approach is often called *digital phenotyping*, entailing the use of smartphones, wearables, and ambient sensors to stream accelerometry, GPS traces, keystroke dynamics, speech acoustics, heart rate variability, and other passively captured metadata. Those streams are preprocessed and feature-engineered [25,26] and then fed into statistical or deep learning models. Parallel work in *affective computing* [27] extends the approach to facial microexpressions, vocal prosody, or text sentiment to classify discrete emotions or arousal levels in real time [28].

As CP systems sit at the intersection of pervasive sensing and advancements in AI, they raise many of the same ethical issues highlighted in broader AI systems. Concerns about algorithmic bias, transparency, explainability, interpretability, fairness, and other aspects of "trustworthy" AI [29,30] are relevant. The rarity with which CP tools are validated on large, diverse validation cohorts means that algorithmic performance is likely to vary dramatically across demographic and clinical groups, raising reliability concerns and potentially amplifying health disparities [31]. Critics have also warned against overreliance on algorithmic inferences about patients' health status [32,33], especially in "black box" systems that resist clinical scrutiny and accountability and compromise informed clinical decision-making [34]. Others [15,35] underscore legal uncertainties surrounding liability in cases of error, patient harm, or mismanagement of outputs or other feedback. The US National Institute of Standards and Technology's AI Risk Management Framework [36] and the European Union AI Act [37] both categorize health-related CP tools as "high risk," demanding rigorous safety, fairness, and oversight provisions.

Similar to other AI systems, CP tools thrive on voluminous datasets, not only across individuals but also for each individual, often referred to as individual "big data" or "deep data" [38]. Ethical critiques thus consistently foreground privacy vulnerabilities associated with sensitive behavioral data [13,14,39,40], and there is expert consensus [13] around the need for privacy and innovative consent approaches. Scholars (eg, [35,41] and C Deeney, BA et al, unpublished data, August 2024) caution against unwanted or involuntary disclosure to third parties, such as insurers or employers, especially in scenarios where data are controlled by consumer-grade device companies. Dynamic consent models have also been proposed [40] to replace onetime or broad consent approaches with ongoing, granular permissions; however, feasibility remains challenging [42].

Challenges for Humanistic Care

Critics [9,43,44] have also converged on a deeper worry: as algorithms assume a larger share of the responsibility to observe and listen, the relational core of care risks being reduced to a "metrics management" exercise, where clinicians and patients spend their limited time consulting data trends rather than discussing the patient's lived experience and therapeutic goals. Clinicians fear that multimodal dashboards could displace narrative dialogue, shifting the burden of self-monitoring and, by extension, responsibility for changes in functioning onto patients in ways that compromise dignity and mutual trust [9,45-47] and overprioritize technological over humanistic solutions [48]. Some warn that automated detection and treatment of illness may weaken the rapport and goal alignment that bolster the therapeutic alliance, unless paired with explicit, empathic communication strategies [49].

A limited set of empirical work reinforces these cautions. One study [47] documented mental health clinician enthusiasm for gaining rich, real-time insights but also highlighted concerns about workflow overload and the potential for automation bias, that is, deferring to algorithmic outputs even when they conflict with a clinician's intuitions or a patient's lived story. Another study [46] highlighted clinicians' concerns that prioritizing passive data trends over self-reported narratives or active responses to clinical assessments could reduce opportunities for patients to reflect on their mental health, leading to diminished patient engagement. Experts [9,21] have raised flags that such asymmetries can tilt encounters toward dehumanization and require careful planning and implementation to achieve the goal of making otherwise invisible patterns visible and clinically useful.

These relational stakes bring long-standing ethical principles into focus and urge clinicians and researchers to keep dignity, empathy, patient empowerment, and shared decision-making at the forefront of clinical care. However, it remains unclear how best to do this in ways that engage multiple and often competing perspectives. Our study addresses this gap by exploring the range of concerns through interviews with over 100 stakeholders who design, deploy, and are the intended users of CP technologies. We catalogue considerations that extend beyond well-elaborated privacy and bias debates to the less operationalized relational harms that data-centric care may impose. By situating these concerns within established humanistic frameworks of dignity, empathy, and shared decision-making [17,18], we offer an anticipatory road map for researchers, developers, clinicians, and patients. The goal is not merely to identify technical fixes, but to ensure that as CP systems mature, they deepen rather than diminish the person-centered relationships that remain the centerpiece of care.

Methods

Study Design

As part of a 4-year study funded by the National Center for Advancing Translational Sciences (R01TR004243), we conducted in-depth, semistructured interviews (total n=102), including adolescent patients (n=20) and caregiver (n=20) dyads, clinicians (n=20), developers (n=21), and ethics, legal, policy, and philosophy scholars (n=21), to explore their perspectives on potential benefits, risks, and concerns around the integration of CP technologies into care.

Participants

Respondents were recruited from a "sister" study (5R01MH125958) aiming to validate CP tools designed to quantify objective digital biobehavioral markers of socioemotional functioning. Participants included a clinical sample of adolescents (aged 12-17 years) with varied diagnoses, including autism, Tourette, anxiety, obsessive-compulsive disorder, and attention-deficit/hyperactivity disorder, as well as their caregivers (typically biological parents; Table 1). Diagnostic presentations for all adolescents were confirmed by expert providers using established clinical measures. Adolescent-caregiver dyads were referred to the study by the sister study's coordinator and then contacted by a research assistant via phone or email to schedule an interview. Clinicians and developers (Table 2) were identified through an online literature search and existing professional networks. Participants were interviewed between January 2023 and August 2023.

Table 1. Demographics for interviewed adolescents and caregivers.[a]

Demographics	Adolescents (n=20), n (%)	Caregivers (n=20), n (%)	Total (N=40), n (%)
Gender			
Male	12 (60)	2 (10)	14 (35)
Female	8 (40)	18 (90)	26 (65)
Race			
American Indian or Alaska Native	0 (0)	1 (5)	1 (3)
Asian	1 (5)	1 (5)	2 (5)
Native Hawaiian or Other Pacific Islander	0 (0)	0 (0)	0 (0)
African American/Black	5 (25)	4 (20)	9 (23)
White	17 (85)	15 (75)	32 (80)
Ethnicity			
Hispanic or Latino	4 (20)	2 (10)	6 (15)
Not Hispanic or Latino	16 (80)	18 (90)	34 (85)
Marital status			
Married and living with spouse	N/A[b]	13 (65)	13 (33)
Widowed	N/A	1 (5)	1 (3)
Divorced	N/A	4 (20)	4 (10)
Separated	N/A	1 (5)	1 (3)
Never Married	N/A	1 (5)	1 (3)
Education level			
High school only or less	N/A	0 (0)	0 (0)
Trade school/associate’s degree	N/A	2 (10)	2 (5)
Bachelor’s degree	N/A	10 (50)	10 (25)
Master’s degree	N/A	4 (20)	4 (10)
Doctoral degree	N/A	4 (20)	4 (10)
Parental status			
Biological parent	N/A	18 (90)	18 (45)
Step parent	N/A	0 (0)	0 (0)
Adoptive parent	N/A	2 (10)	2 (5)
Diagnosed condition			
Obsessive-compulsive disorder	4 (20)	N/A	4 (10)
Autism	5 (25)	N/A	5 (13)
Attention-deficit/hyperactivity disorder	3 (15)	N/A	3 (8)
Anxiety	4[c] (20)	N/A	4 (10)
Tourette	1 (5)	N/A	1 (3)
No clinical diagnosis or symptoms	9 (45)	N/A	9 (23)
Average age, mean (SD)	14.9 (2.2)	48.3 (6.4)	N/A

[a]Values may not total 100% owing to overlapping categories (eg, comorbidities), nonmutually exclusive response options, and skipped questions.
[b]N/A: not applicable.
[c]1 self-reported.

Table 2. Demographics for interviewed clinicians, developers, and ELPP.[a]

Demographics	Clinicians (n=20), n (%)	Developers (n=21), n (%)	Scholars (n=21), n (%)	Total (N=62), n (%)
Gender				
Male	10 (50)	18 (86)	16 (76)	44 (55)
Female	10 (50)	3 (14)	5 (24)	18 (29)
Race				
American Indian or Alaska Native	0 (0)	0 (0)	0 (0)	0 (0)
Asian	4 (20)	1 (5)	2 (10)	7 (11)
Native Hawaiian, Pacific Islander, or Other	0 (0)	0 (0)	0 (0)	0 (0)
African American/Black	0 (0)	1 (5)	0 (0)	1 (2)
White	14 (70)	12 (57)	16 (76)	42 (68)
Unreported/unknown	3 (15)	6 (29)	4 (19)	13 (21)
Ethnicity				
Hispanic or Latino	0 (0)	0 (0)	1 (5)	1 (2)
Not Hispanic or Latino	17 (85)	13 (62)	16 (76)	46 (74)
Unreported/unknown	3 (15)	7 (33)	4 (19)	14 (23)
Profession				
Clinician	3 (15)	N/A[b]	N/A	3 (5)
Clinician-researcher	14 (70)	N/A	N/A	14 (23)
Clinician-developer	3 (15)	4 (19)	N/A	7 (11)
Developer	N/A	17 (81)	N/A	17 (34)
Ethicist	N/A	N/A	6 (29)	6 (10)
Lawyer	N/A	N/A	4 (19)	4 (6)
Philosopher	N/A	N/A	1 (5)	1 (2)
Other	N/A	N/A	10 (48)	10 (16)
Specialty				
Psychiatry	7 (35)	N/A	N/A	7 (11)
Psychology	7 (35)	N/A	N/A	7 (11)
Neuroscience	4 (20)	N/A	N/A	4 (6)
Industry	N/A	15 (71)	N/A	15 (24)
Academic	N/A	3 (14)	N/A	3 (5)
Cross-Sector	N/A	3 (14)	N/A	3 (5)
Ethics	N/A	N/A	6 (29)	6 (10)
Law	N/A	N/A	4 (19)	4 (6)
Philosophy	N/A	N/A	1 (5)	1 (2)
Other	2 (10)	N/A	10 (48)	12 (16)

[a]Some values may not total the number of stakeholders per group or 100% because certain responses were missing, some response options were nonmutually exclusive, and respondents were allowed to skip questions.

[b]N/A: not applicable.

Data Collection

Separate but parallel interview guides were developed for all stakeholders, with the same constructs explored across groups, including perceived benefits and concerns regarding integrating CP tools into clinical care; impacts on care; attitudes toward automatic and passive detection of emotional and behavioral states; perceived accuracy and potential for misinterpretation, misattribution, or misclassification of symptoms or conditions; clinical utility and actionability; data security and privacy concerns; potential for unintended uses; perceived generalizability and potential for bias; and other emergent concerns. These domains were chosen based on issues raised in the clinical and ethics literature and with guidance from experienced bioethicists and mental health experts. Initial drafts of the interview guides were piloted with 2 psychologists (EAS and CJZ) specializing in adolescent mental health, resulting in minor clarifications in wording. Interviews were conducted via a secure videoconferencing platform (Zoom for Healthcare; Zoom Communications, Inc) and lasted an average of ~45 minutes. Participants watched a brief 1.5-minute "explainer" video defining CP as denoting AI systems (devices + algorithms) that not only sense but also infer and act upon multimodal behavioral and physiological signals. Demographic items were included to explore possible sociodemographic variation in perspectives and to facilitate downstream analyses or meta-analytic comparisons. Participants could select more than 1 racial or ethnic category, and no participant was required to respond to any demographic question.

Ethics Approval

This study was reviewed and approved by the Baylor College of Medicine Institutional Review Board (H-52227), which waived the requirement for written consent, as the research procedures (interviews, deidentification of transcripts, and storage on secure servers) involved minimal risk to participating stakeholders; thus, participants provided verbal consent. Minors provided assent with parental consent. Identifiable participant information was stored behind a university firewall in a password-protected system with 2-factor authentication. All results are reported in aggregate and not linked to any identifiable participants, including in supplementary documents. All participants also completed a brief demographic questionnaire in REDCap (Research Electronic Data Capture; Vanderbilt University) via an emailed link.

Data Analysis

Interviews were audio-recorded, transcribed verbatim, and analyzed using MAXQDA software (VERBI Software). Led by a qualitative methods expert (KMKQ), team members developed a codebook to identify thematic patterns in adolescent and caregiver responses to the topics described above. Each interview was coded by merging the work of 2 separate coders to reduce interpretability bias and enhance reliability. All team members received extensive training in qualitative analysis before participating in coding. We used thematic content analysis [50,51] to inductively identify themes by progressively abstracting relevant quotes, a process that entails reading every quotation to which a given code was attributed, paraphrasing each quotation (primary abstraction), further identifying which constructs were addressed by each quotation (secondary abstraction), and organizing constructs into themes. The multidisciplinary team responsible for thematic analysis consisted of the principal investigator (KMKQ), who is a medical anthropologist and bioethicist with expertise in qualitative and mixed methods research, bioethics, and the social and ethical dimensions of AI and digital phenotyping; and 3 research associates—2 master's-level researchers with backgrounds in psychology, neuroscience, bioethics, and cognitive science, and 1 postbaccalaureate researcher with training in psychology and computer science. This combination of disciplinary and methodological perspectives was intentionally designed to reduce interpretive homogeneity and promote reflexivity. To enhance the validity of our findings, all abstractions were validated by at least one other member of the research team. In rare cases where abstractions reflected different interpretations, members of the research team met to reach consensus. Coding meetings emphasized interpretive dialogue rather than consensus by conformity, ensuring that thematic reliability reflected triangulation across diverse epistemic standpoints rather than agreement among individuals with similar expectations. Frequencies were also calculated for each theme by counting the number of individuals within each stakeholder group who contributed at least one quote coded under that theme. These frequencies and percentages are presented solely as descriptive indicators and are not intended to imply statistical significance or support inferential claims.

Results

Themes Identified

Stakeholders raised a wide range of concerns around the following themes (Table 3): (1) Trustworthiness and Integrity of CP Technologies (72/102, 70.6%; Multimedia Appendix 1); 2) Patient-Specific Relevance (24/102, 23.5%; Multimedia Appendix 2); (3) Utility and Implementation Challenges (60/102, 58.8%; Multimedia Appendix 3); (4) Regulation and Governance (17/102, 16.7%; Multimedia Appendix 4; (5) Data Privacy and Protection (88/102, 86.3%; Multimedia Appendix 5); (6) Patient Harms (65/102, 63.7%; Multimedia Appendix 6); and (7) Philosophical Critiques (13/102, 12.7%; Multimedia Appendix 7). All themes and subthemes are elaborated below, with illustrative quotations in the associated Multimedia Appendices 1-7.

Table 3. Theme frequencies.[a]

Theme	Developers (n=21), n (%)	Clinicians (n=20), n (%)	Adolescents (n=20), n (%)	Caregivers (n=20), n (%)	Ethics, law, policy, and philosophy scholars (n=21), n (%)	Total (N=102), n (%)
Trustworthiness and Integrity	17 (81)	15 (75)	8 (40)	15 (75)	17 (81)	72 (70.6)
Patient-Specific Relevance	3 (14)	3 (15)	3 (15)	7 (35)	8 (38)	24 (23.5)
Utility and Implementation	15 (71)	15 (75)	4 (20)	9 (45)	17 (81)	60 (58.8)
Regulation and Governance	4 (19)	4 (20)	0 (0)	1 (5)	8 (38)	17 (16.7)
Data Privacy and Protection	16 (76)	17 (85)	16 (80)	20 (100)	19 (90)	88 (86.3)
Patient Harms	9 (43)	16 (80)	4 (20)	18 (90)	18 (86)	65 (63.7)
Philosophical Critiques	2 (10)	2 (10)	0 (0)	2 (10)	7 (33)	13 (12.7)

[a]Frequencies and percentages are calculated within groups except for when they are in the "Total" column, where they are calculated across groups.

Trustworthiness and Integrity of CP Technologies

Data Quality Constraints and Confounds

Developers, more than other stakeholder groups, raised concerns about the reliability of data streams from consumer-grade devices, emphasizing that variations in user behavior and differences in hardware performance can make it difficult to distinguish true physiological changes from sensor-related errors. They cautioned that without standardized protocols for device calibration and data collection, models built on such inputs may fail when deployed across different environments or patient populations.

Algorithmic Bias and Generalizability

Participants across all stakeholder groups also raised concerns about additional forms of algorithmic bias. Several scholars noted that many AI models are trained on relatively homogenous datasets, limiting their generalizability to more diverse populations. As these datasets often disproportionately represent individuals from more privileged groups (eg, younger, healthier, or majority-ethnic cohorts), algorithms may underperform or misclassify signals in marginalized communities. Participants further cautioned that unequal access to digital health technologies can skew training data even more, reinforcing systemic biases and potentially excluding the very populations most likely to benefit from improved care.

Construct Validity

Clinicians, developers, and scholars alike cautioned that the diagnostic constructs and clinical assessment tools used to validate most CP tools often lack strong links to clinically meaningful phenomena and fail to accommodate transdiagnostic symptom presentations, cultural and contextual variability, and temporal fluctuations in mental health. As a result, the digital markers derived from these tools risk remaining insufficiently grounded. Participants emphasized the need for rigorous validation studies to ensure that digital biomarkers accurately reflect patient states and that any interventions based on these measures are anchored in well-established clinical evidence.

Patient-Specific Relevance

Accounting for Heterogeneity in Symptom Expression and Subjectivity

Stakeholders consistently emphasized that any use of digital health tools must first account for the immense diversity in how individuals experience and express their health and then situate those signals within each person's unique context. Respondents across groups cautioned that a one-size-fits-all algorithm may miss or misinterpret patients who exhibit emotional or behavioral states differently from others; for example, some noted that while certain individuals express distress outwardly, others internalize such feelings, rendering them "invisible" to CP tools searching for external markers. Others added that accurate interpretation often depends on integrating multiple data streams; heart rate alone, for instance, may not distinguish stress from exercise without information about the broader context or behavioral pattern.

Accounting for Context and Meaning

Patients and caregivers, more than other groups, raised concerns that algorithms cannot effectively account for the rich social and cultural factors that shape patients' experiences and behaviors, or how patients assign meaning to their symptoms and events. Some also emphasized the importance of proximate contextual features, such as fluctuations tied to work demands, family stressors, or lifestyle changes. Patients, in particular, worried that algorithms might draw conclusions based on fleeting or temporary signals rather than longer-term trends. Respondents across groups cautioned that such "decontextualized" metrics lack the construct validity required for clinical actionability, as they are likely to reflect inferences stripped of subjective meaning and, therefore, clinical significance.

Utility and Implementation Challenges

Role of CP in Clinical Care

Stakeholders from all groups voiced a set of interrelated concerns about how CP tools are integrated into clinical workflows. Scholars and clinicians cautioned that clinicians may lean too heavily on algorithmic outputs, risking a form of "deskilling" in which they stop rigorously scrutinizing the data for quality or epistemic inconsistencies. They warned that clinicians may begin to accept CP suggestions uncritically (automation bias), thereby sidelining the human, relational interpretations developed through patient-provider dialogue.

Managing Risk and Liability

Clinicians, more than other groups, highlighted the dual dangers of missed events and overalerting. They noted that false negatives—instances where the system fails to detect deterioration—could leave patients unprotected, while excessive false positives could overwhelm clinicians and erode confidence in the tool, ultimately undermining patient safety rather than enhancing it. Clinicians also raised concerns about whether they may eventually be expected to use CP tools as these systems continue to evolve, or held liable if they choose not to, thereby compromising their autonomy in clinical decision-making.

Barriers to Utility

All stakeholder groups stressed that CP outputs must be interpretable and meaningful in real-world contexts to be actionable. Clinicians emphasized that data trends and inferences should be delivered through intuitive summaries and visualizations, accompanied by concise, actionable recommendations. They noted that this is complicated by the fact that the clinical significance of data trends may vary from one situation to another (see the "Patient-Specific Relevance" section), making consistent interpretation challenging. Developers and clinicians also raised concerns about the potential for confirmation bias, in which users may selectively interpret or emphasize data that confirm their expectations, thereby undermining the goal of these technologies to contribute novel informational value to clinical assessments.

Regulation and Governance

Unclear or Insufficient Regulatory Frameworks

Clinicians and scholars, more than other groups, described 2 distinct but related regulatory challenges. First, many CP applications can (and in their view, should) fall under existing clinical-use regulations, such as those governing medical devices; yet, few concrete guidelines exist for implementing these requirements in practice. Ethics and policy experts noted that when CP tools nominally qualify as regulated devices, organizations may feel more comfortable adopting them; however, the absence of clear, step-by-step governance pathways often leaves developers and clinicians uncertain about how to operationalize data privacy, security, and ethical review processes. Second, participants emphasized that a large swath of CP technologies occupies a "regulatory gray zone" due to their overlap with devices in the consumer "wellness" sector, particularly those that collect passive or contextual data outside traditional care encounters. Scholars worried that without specifying oversight requirements for continuous, ambient monitoring, regulators risk leaving patients exposed to unvetted algorithms and unclear lines of accountability.

Responsibility for Ethical Technology Development and Compliance

Developers, scholars, and clinicians primarily expressed concerns about how innovation pressures interact with ethical safeguards. On the one hand, experts described the burden of balancing innovation against regulatory demands, noting that small teams sometimes struggle to absorb the time and cost required for formal ethics and security reviews. They also raised concerns about the deployment of closed-source, proprietary algorithms, which are often faster to market but opaque. These were contrasted with open-source alternatives, which permit external audit but come with greater technical support obligations. Across both choices, questions about liability were raised, with respondents arguing that without explicit legal clarity, neither developers nor health care providers know with certainty who would be held accountable if CP assessments lead to harm.

Need for Stakeholder Involvement

Respondents from all groups expressed strong consensus that regulation and governance structures must be co-designed with the people intended to benefit from these technologies. Ethics scholars argued that embedding patients' and caregivers' lived experiences into standards setting is vital to ensuring that tools address real-world needs. Clinicians highlighted the importance of rigorously interrogating when and under what circumstances CP outputs truly matter to patient care, rather than assuming that technological assessments will always be relevant. Participants across groups also called for interdisciplinary collaboration among technologists, clinicians, ethicists, and end users to bridge gaps in expertise, surface hidden risks, and develop governance models that are both practical and ethically robust.

Data Privacy and Protection

Consent and Awareness

Patients described anxiety about unwanted or unintended disclosure of intimate behavioral and physiological data, noting that continuous collection can feel like a privacy breach. Other stakeholder types likewise questioned the appropriateness of capturing real-time location or mental health indicators, characterizing such practices as invasive and, in some cases, "creepy." This unease was compounded by awareness that elements of coercion may come into play: individuals could feel pressured to share their data so as not to jeopardize access to health care services. Adding to these worries, stakeholders noted that explanations of data practices are often obscure, leaving patients unaware or uncertain about what exactly they are consenting to, who may access their data, what inferences could be drawn from it, and what kinds of feedback to expect. As a result, patients may be ill-equipped to make informed decisions about engaging with these CP tools or about what types of feedback to receive or decline (eg, exercising a "right not to know").

Many participants, especially researchers, clinicians, and ethics scholars, criticized current informed consent practices as outdated and one-dimensional. They noted that patients typically encounter a single form at the outset of care (broad consent) without fully understanding the breadth of data being collected or the myriad ways it might later be used. Several respondents urged a shift toward dynamic consent models, in which patients receive clear, ongoing explanations and can granularly and dynamically opt in or out of specific data uses. They emphasized that such processes—which treat consent as an evolving conversation—are better suited to the continuous, ecological monitoring characteristic of CP approaches.

Secondary Use and Misuses

Many patients and caregivers reported being comfortable with primary clinical uses of CP data but expressed concern about secondary applications and potential misuses. Stakeholders across groups noted that, without clear legal protections, patient information could be repurposed for discriminatory profiling or accessed by commercial actors, with existing regulations offering little guidance on how to manage these downstream uses. They argued that the commodification and monetization of personal behavioral and physiological data, in the absence of robust data protection frameworks, could erode patient and caregiver trust in clinicians and health care systems and discourage future participation in digital health programs.

Monitoring and Surveillance

Stakeholders also observed that when individuals feel monitored rather than supported, they may withhold information, worry about data misuse, and question their providers' trustworthiness. This concern may be particularly relevant for vulnerable populations, such as people experiencing psychosis, who may perceive passive monitoring as surveillance, and older adults who may have difficulty using wearables and apps—highlighting the need for adaptive protocols, additional safeguards, and alternative engagement strategies that respect each patient's autonomy and comfort. They emphasized that passive monitoring can shift the experience from feeling supported to feeling observed, an effect that may be especially pronounced among vulnerable groups; for example, people experiencing psychosis may interpret continuous tracking as intrusive surveillance, and members of historically exploited populations may hold significant reservations.

Patient Impacts and Harms

Overview

Stakeholders highlighted numerous ways in which the above concerns may translate into direct or indirect harms for patients:

Harms Due to Inaccurate or Premature Diagnoses

Stakeholders from all groups cautioned that algorithmic assessments delivered without sufficient clinical context can trigger a cascade of inappropriate interventions. They warned that acting on false positives or early "flags" could expose patients to unnecessary tests, treatments, or stigma long before a human expert has had an opportunity to validate the finding. They also noted the potential negative impacts when algorithmic conclusions diverge from patients' own perceptions and experiences, creating conflict without clear pathways for resolution.

Diminished Human Connection in Health Care

A recurring theme, particularly among clinicians and patients, was the potential breakdown of the human connection in health care. Many stakeholders noted that an overreliance on data-driven CP tools could transform care into a more transactional and less empathetic process. Clinicians especially underscored the importance of maintaining therapeutic relationships grounded in respect, empathy, and alliance, warning that digital tools—while potentially efficient—could diminish the "human touch" that is central to healing. Many patients and caregivers echoed this concern, fearing that health care interactions could become increasingly impersonal. Scholars and clinicians also discussed the potential for digital health tools to contribute to epistemic injustice, whereby patients' lived experiences may be undervalued in comparison to data-driven assessments. Some stakeholders expressed concern that an emphasis on objective data could lead clinicians to discount patients' subjective experiences—especially in complex domains such as mental health, where self-reports already face considerable scrutiny. They warned that such dismissal could erode patient autonomy and contribute to a dehumanization of care, particularly if clinicians and patients allow algorithmic inferences to assume an increasingly prominent role relative to human judgment in decision-making.

Responsibility Shifts and "Empowerment" Pitfalls

Another significant concern raised by clinicians involved the shifting of responsibility from health care providers to patients. As digital tools increasingly monitor and manage health, patients are often expected to assume a larger role in their own care. While some viewed this shift as empowering, many clinicians feared it could overwhelm patients—especially those without the skills, knowledge, or interest to interpret continuous data feedback—potentially leading to confusion, stress, and unintended burdens.

Ethics scholars also noted that although the rhetoric of "empowerment" is often used to promote these tools, it can effectively shift responsibility onto individuals—particularly those with greater resources—while leaving vulnerable populations with few mechanisms to address complex health inequalities. They emphasized that this shift not only places an undue burden on patients to manage their health independently but also predisposes them to blame when improvements do not occur, potentially worsening feelings of shame or anxiety. Several ethics and policy scholars argued that this trend is reinforced by the technology sector's tendency to view patients as consumers rather than individuals needing care, thereby framing health management as an individual rather than a collective responsibility.

Additionally, clinicians noted the risk that patients may defer responsibility to technology—such as smartphones—under the assumption that these tools will manage their health for them, which can diminish active engagement in their own care. They cautioned that when patients come to believe that their devices will "speak" on their behalf, they may become less inclined, and over time less able, to reflect on and articulate their own experiences and behavioral patterns.

Access Inequities and Disproportionate Burdens to Vulnerable Populations

Clinicians and scholars voiced further concerns about the potential of CP tools to exacerbate inequities and disproportionately burden vulnerable populations. Scholars emphasized that marginalized groups—including those experiencing poverty, homelessness, and other forms of marginalization—may be excluded from the benefits of these technologies due to a lack of access or capacity. For example, individuals without consistent access to, or familiarity with, technology might struggle to effectively use or trust these tools, limiting potential benefits and skewing training datasets in ways that perpetuate harmful biases and further exacerbate inequities.

Further, caregivers and ethicists, in particular, raised significant reservations about CP tools being leveraged or co-opted for surveillance, especially in communities with a history of being monitored, such as psychiatric and other vulnerable groups. Pressured consent emerged as another concern, particularly for individuals in lower social positions who might feel compelled to use these tools despite discomfort or uncertainty. Finally, stakeholders highlighted the risk of involuntary monitoring or detention, noting that misdiagnoses or inaccurate data could lead to wrongful decisions with severe consequences for individuals' rights and treatment.

Threats to Privacy and Self-Determination

Stakeholders from all groups voiced strong concerns about the threats to privacy and autonomy posed by digital health tools. They highlighted the potential misuse of sensitive health data and the lack of transparency in how such information is collected and used. Scholars emphasized the need for stronger regulatory frameworks to ensure that patients' privacy is protected and that they retain control over their personal health data. They warned that without adequate safeguards, the widespread adoption of these technologies could lead to breaches of trust and unauthorized access to sensitive information.

Clinicians noted that certain patient populations are likely to be disproportionately affected by these concerns and may require particularly robust clinical justifications, as well as enhanced protections or alternative approaches, to ensure that CP tools benefit their care while safeguarding their rights to self-determination and protection against discrimination.

Epistemic Injustice and Deprioritization of Patient Voices

Stakeholders cautioned that CP tools risk sidelining patients' own experiences by privileging algorithmic inferences over first-person testimony. Ethics scholars noted that even highly accurate systems can produce outputs that contradict a patient's self-knowledge, potentially leading clinicians to discount lived perceptions and destabilize trust. Caregivers emphasized that real-time observations—such as a parent's instinct about a child's well-being—must carry equal or greater weight than sensor data to avoid silencing those closest to the patient.

Overemphasis on Self-Optimization

Experts warned that voluntary self-tracking can evolve into a cultural expectation, similar to how smartphones have become indispensable. What begins as clinically guided monitoring risks morphing into relentless personal optimization, pressuring individuals to engage in continuous self-surveillance. Stakeholders argued that blurring the line between medical indication and consumer-driven tracking reduces complex human experiences to mere data points and undermines broader notions of well-being that cannot be quantified.

Philosophical Critiques of CP

CP Is Insufficient to Capture Emotional States

Certain scholars cautioned that CP technologies cannot fully capture the rich complexity of human emotion. They argued that feelings are not reducible to physiological impulses or static signals, but instead unfold in nuanced, dynamic patterns that resist algorithmic measurement.

CP Cannot Infer Emotion via Behavior

Relatedly, some stakeholders emphasized that CP tools cannot reliably infer emotion from behavior alone. While sensors can record facial movements, voice acoustics, heart rate fluctuations, and other behavioral or physiological signals, these outward markers do not necessarily reflect internal experience and always require human interpretation. One scholar likened this need for interpretation to how a radiologist must analyze and contextualize an image.

CP Algorithms Embed Human Biases

Other participants emphasized that, because CP algorithms inevitably incorporate human biases, they cannot serve as purely objective indicators of pathology. They noted that every algorithm is trained on manually labeled data and thus carries forward the cultural assumptions and biases of its creators. They argued that reliance on precoded categories can obscure these underlying prejudices by presenting CP outputs as seemingly "objective."

CP Inferences Are Not More Valuable Than Subjective Patient Insights

Some scholars challenged the overprioritization of data over dialogue, emphasizing that personal narratives—rooted in lived, phenomenological experience—provide primary and indispensable insights into illness that digital metrics cannot replace. They contended that patient testimony must "stand on equal footing" with any algorithmic outputs.

CP Reflects Techno-Solutionism

Scholars warned that addressing illness primarily through a technological lens reflects a broader misconception that technology can solve all problems. They emphasized the importance of attending to the social, political, and cultural dimensions of health. These stakeholders argued that an overemphasis on what can be measured or automated risks shaping health care interventions around the capabilities of machines rather than the holistic needs of people.

Discussion

Corroborating Existing Recommendations

Our investigation highlights the broad and varied concerns of diverse stakeholders—developers, clinicians, patients, caregivers, and ethics and policy experts—regarding the integration of CP into clinical care. Understanding and addressing these concerns is critical for designing implementation strategies that enhance, rather than compromise, patient-centered and humanistic care. Many of the themes echo longstanding critiques of data-centrism in medicine: CP represents the latest iteration of placing ever richer "deep data" streams at the center of care, now amplified by powerful AI and machine learning analytics. Accordingly, stakeholders reiterate familiar principles from the trustworthy AI framework, including explainability, interpretability, bias mitigation, fairness, and transparency. The opaque, "black-box" nature of many proprietary CP algorithms further compounds these challenges, leaving patients and caregivers without clear evidence of how inferences about mood, cognition, or behavior are generated. Respondents in our study, echoing prior calls, advocate for robust, domain-specific validation standards, enhanced algorithmic transparency, liability frameworks for errors, mechanisms for contesting outputs, and guidance on reliably interpreting CP results across diverse clinical settings. These imperatives are neither new nor contested; there is a broad consensus on the need for trustworthy algorithms coupled with humanistic care.

Similarly, the call for implementation frameworks that protect clinician judgment, patient agency, and the therapeutic alliance is well established. Stakeholders cautioned that uncritical, algorithm-driven monitoring risks displacing empathic dialogue by prioritizing decontextualized or biased metrics over patients' own narratives, shifting the therapeutic focus from shared understanding to automated inference. These concerns are most pronounced for CP systems that directly infer diagnosis (classification) or prognosis (prediction), but may be less significant when CP is used to surface raw patterns—such as sleep or activity metrics—for human-guided interpretation. For example, rather than allowing an algorithm to label sleep patterns as pathological, clinicians could use a patient's baseline sleep data—compared with population benchmarks—to ask, "What's keeping you up at night?" and collaboratively determine what constitutes normal sleep for that individual in the context of work, family, or lifestyle factors. D'Alfonso and colleagues [9] describe this distinction as "manual" versus "AI-driven" use of CP, emphasizing the degree of human involvement in interpreting data. At the time of writing, most CP tools are not yet robust enough to rely solely on AI-driven inferences and therefore require substantial human interpretation to be clinically useful. However, as we argue elsewhere [52], this may not always remain the case; following the trajectory of AI in other domains, CP algorithms are likely to evolve to provide valid, accurate, patient-specific, and trustworthy inferences. Establishing humanistic approaches well in advance is a widely recognized and consensus goal.

Novel Insights: The Importance of Context and Subjectivity

Our respondents highlighted 2 fundamental considerations for effectively and humanely integrating CP tools into care that have not been fully addressed elsewhere: the importance of context and subjectivity in determining the clinical significance of CP outputs. Stakeholders across all groups emphasized that observable behaviors—such as steps, voice tone, and facial micro-movements—are clinically actionable only when clinicians understand what those behaviors signify for the individual producing them and how the surrounding context shapes that meaning.

This caution echoes the "Theory of Constructed Emotion" proposed by Barrett et al [53] and supported by like-minded scholars [54-57], who challenge the classical view that emotions are biologically hard-wired states expressed through universal behavioral markers. Instead, the brain constructs each feeling from past experiences, cultural learning, and moment-to-moment interpretation; the same smile, for example, can signify joy, embarrassment, or compliance depending on context [28,58]. When CP systems infer affect solely from facial features, vocal prosody, heart rate variability, or other external cues, they risk reducing this complexity to generic labels—an error that disproportionately misinterprets individuals across different cultures, age groups, or clinical presentations [59].

To counter such reductionism, future CP strategies must integrate subjective meaning and environmental context alongside sensor data. Technically, this involves pairing passive streams with structured self-report or ecological annotations that capture the patient's interpretation of events and the situational factors influencing them. Operationally, it requires structured conversations—from the earliest visits through follow-up—that identify which symptoms most constrain a person's quality of life and how those symptoms might be detected digitally. The "Digital Measures That Matter to Patients" framework proposed by Manta and colleagues [60] provides concrete guidance, linking *meaningful aspects of health* to sensor-derived concepts of interest, outcomes, and end points within a patient-centered hierarchy.

In practice, applying this framework could mean, for example, that a patient who values uninterrupted sleep over daytime mood stability prioritizes actigraphy-based sleep metrics, whereas another concerned about social withdrawal might ask the system to flag sustained reductions in communication patterns. By integrating patient narratives and contextual details into metric selection and interpretation, clinicians can transform CP from a one-size-fits-all detector into a context-aware, individually tailored decision-support tool—remaining faithful to the subjective richness that stakeholders emphasize must never be lost.

A Prototype for Humanistic Care With CP

Personalized Road Maps for CP Integration

To address these challenges, we introduce the concept of personalized road maps [61] for integrating CP into clinical care—a structured, co-designed plan that embeds humanistic values at every stage of digital phenotyping. Rather than treating data feedback as a series of discrete disclosures, personalized road maps are collaboratively developed by patients, caregivers, and clinician-researchers at the point of consent. Together, they specify the following:

- *Which metrics* (eg, activity patterns, speech markers, sleep variability) will be monitored and shared
- *When and how* these data will be returned—whether in real time, during clinic visits, through periodic summaries, or some strategic (nonarbitrary) mix of approaches
- *Thresholds for action,* delineating what combinations of signals should trigger outreach, referral, or adjustment of treatment

- *Conflict resolution procedures* for managing epistemic conflicts when CP outputs diverge from a patient's self-report or a clinician's judgment.

This iterative framework balances patient agency with clinical and ethical guardrails, inviting patients to contribute lived knowledge (eg, recognizing that reduced SMS text messaging often precedes mood dips), while researchers share their clinical expertise. Together, both parties anticipate and develop shared understandings of how their perspectives may be enriched by predictive insights from CP data trends and inferences. This approach reflects a view, articulated by others [49], that technology and humane care are not mutually exclusive, but can, in fact, be symbiotic. The personalized road map is designed to foster that symbiosis, serving as a living decision-support tool that aligns computational power with at least three operationalized, person-centered goals of care, including those listed below.

Empowerment and Shared Decision-Making

By inviting patients to coselect which CP signals matter most and how they wish to receive feedback, personalized road maps transform passive monitoring into an active partnership. This builds on Schmidt and D'Alfonso's [47] finding that clinicians and clients value systems where patients can "switch off" sensors, control data sharing, and iteratively refine monitoring parameters. Patients can collaboratively choreograph the timing, dose, and content of feedback to align with their treatment goals. Embedding these choices upstream helps prevent downstream surprises or distress when digital inferences arise.

Trust and Therapeutic Alliance

Clear, cocrafted expectations—about what data will be returned, when, and under what conditions—help mitigate nocebo effects and overreliance on opaque risk scores. As Nghiem et al [46] observed, passive patient-generated health data are most useful when presented at clinically meaningful moments, rather than overwhelming clinicians in real time. Personalized road maps can specify this timing, ensuring that data review occurs within empathetic, dialogic encounters rather than disrupting them.

Ethical Transparency and Anticipation of Conflict

Documenting both the inclusion and exclusion of specific CP metrics is inspired by the "open notes" movement, providing patients with insight into the analytic process. This approach preserves their right to understand which factors shape their treatment pathways, as well as their right "not to know" certain inferences that might be counterproductive to clinical progress. Road maps also embed anticipatory strategies for epistemic conflicts. For example, if a wearable flags elevated stress while a patient reports feeling calm, the road map can offer coidentified strategies to guide the clinician and patient through a respectful dialogue about potential device errors, contextual factors, or unrecognized symptoms, rather than defaulting to algorithmic authority or privileging patient report.

Innovating Consent for CP Approaches

As CP technologies transition from clinical research into routine care, these road maps will support clinical teams in their fiduciary responsibility to educate patients about anticipated benefits and risks, while transparently conveying areas of uncertainty. Enhancing existing consent procedures should begin with identifying the knowledge needs of patients and caregivers to enable truly informed consent. In a recent publication, we reported the results of an empirical, qualitative analysis [62] exploring the perspectives of adolescent patients and their caregivers participating in clinical laboratory research involving extensive CP data collection. Our findings demonstrated that patients and caregivers have information needs spanning 7 key themes: (1) clinical utility and value; (2) evidence, explainability, evaluation, and contestation; (3) accuracy and trustworthiness; (4) data security, privacy, and potential misuse; (5) patient consent, control, and autonomy; (6) the physician-patient relationship; and (7) patient safety, well-being, and dignity. A separate analysis (C Deeney, BA et al, unpublished data, August 2024) found that most patients and caregivers consider CP data highly sensitive and are reluctant to share these data beyond their clinical teams. While many participants expressed trust in existing data protections to safeguard CP data, they often misunderstood or overestimated the extent to which protections such as the Health Insurance Portability and Accountability Act (HIPAA) apply. Based on these findings, we proposed 5 key strategies: (1) educating patients on the limitations of existing data protections; (2) conducting targeted research, including forensic analyses, into secondary data exchanges to identify privacy breaches or reidentification risks; (3) enacting regulations that mandate greater transparency in health data transactions; (4) implementing computational mechanisms, such as distributed ledger technologies, to enhance data traceability and auditability; and (5) adopting dynamic consent models that allow patients to continuously manage and update their consent preferences.

Other scholars have similarly argued that static, onetime signatures are inadequate for the continuous, highly contextual data streams generated by CP tools. A systematic review of ethical considerations for passive data sensing [63] proposed interactive informed consent interfaces that allow participants to add social annotations, "talkback" questions, and multimodal visual aids—features shown to enhance comprehension and engagement [64,65]. Others have called for context-sensitive consent models [66,67], allowing patients to recalibrate permissions as circumstances change and enabling built-in data expiration options, so individuals can set automatic sunset dates [68]. These consent innovations should be embedded within the personalized road map architecture to ensure that consent remains an evolving, rather than static, agreement.

Operationalizing Humanistic Use of CP

Most would agree that maintaining a sense of humanity in care is critical—and in fact, we already have a reasonably clear vision of what humanistic practice entails, even if current systems fall short. Humanistic care is compassionate, respectful, and empathetic. It is also collaborative, culturally sensitive, and empowering. The formative research presented here corroborates a substantial body of prior work [69-71] demonstrating how diverse stakeholders conceptualize and idealize humanistic care. In other words, further studies to delineate what constitutes humanistic practice and to demonstrate its benefits for patients, clinicians, and communities are no longer the priority; that foundational work has already been done. What is now required is rigorous, context-specific evidence identifying which CP integration strategies most effectively embody these established humanistic care ideals—that is, which organizational policies, device design features, relational practices, and value-based attitudes to incorporate, and which to eschew. We still lack evidence-based guidelines for integrating CP, and the only way to develop them is to investigate a wide spectrum of implementation contexts to determine which combinations of features produce desired outcomes, for which patients, and under what circumstances. Our analysis highlights several feature domains that require systematic evaluation:

- *Data handling:* collection methods, governance structures, and privacy safeguards
- *Feedback logistics:* cadence, routing, and escalation pathways
- *Patient support:* education, engagement, and shared-decision tools
- *Analytics:* modeling choices, interpretive aids, and decision-support mechanisms
- usability, accessibility, and visualization elements
- *Workflow integration:* infrastructure requirements and task allocation
- *Clinician readiness:* training, supervision, and capacity building

Each domain contains multiple variables whose effects may differ by setting. Treating these variables as elements of a "constellation" and iteratively testing how their configurations influence clinical and humanistic outcomes will allow us to identify the scenarios in which specific approaches add value—and those in which they do not. Such empirical investigation may reveal that CP approaches are not suitable for every patient or clinical scenario.

Concluding Reflections

Integrating CP technologies into everyday clinical workflows surfaces specific tensions that can undermine even the most deeply held humanistic ideals. Numerous forces compete with our ability—or even our desire—to deliver humanistic care. In the case of CP, one of the most pervasive is the shared conviction—among clinicians, patients, and caregivers alike—that data speak more objectively than lived experience. As our stakeholders cautioned, centering illness interpretations on digital signals risks reframing patients' stories through the lens of machine-generated feedback. Anthropologists describe this phenomenon as an "idiom": a culturally patterned mode of expression—verbal, behavioral, or somatic—through which distress or well-being is communicated in ways that reflect shared meaning based on local beliefs and values. Classic idioms of distress, such as "heavy heart" [72,73], "ataque de nervios" [74], or notions of hot-cold imbalance [75], function less as discrete biomedical signs and more as symbolic languages linking individual suffering to broader cultural meanings, social relationships, and moral concerns. If data become the dominant idiom through which we express or even conceptualize illness, we may lose the ability to recognize, convey, and intervene in the complex multitude of factors influencing health and illness.

These idiomatic shifts pose far graver threats than concerns about false alarms, opaque metrics, or data privacy—issues that, while critically important, are largely tractable and already receiving extensive scholarly and technical attention. By contrast, the greater danger lies in narrowing our collective capacity to perceive human realities by privileging quantifiable signals over the nuanced psychosocial factors that shape how illness is understood and experienced. From this perspective, dehumanized care represents not merely a violation of respect or rights, but a siphoning of human insight, potentially eroding clinicians' curiosity and compassion as well as patients' ability to articulate their own experiences.

Ironically, this outcome runs counter to CP's original promise: to provide objective, reliable insights into complex disease states and, in doing so, bring us closer to the ground truths of human suffering. Data alone cannot constitute those truths. The critical question—one that our study helps illuminate—is how to integrate these deep data into care in ways that strengthen, rather than undermine, the humanistic foundations of clinical practice.

Acknowledgments

We sincerely thank all the clinicians, developers, ELPP scholars, patients, and caregivers who participated in interviews and shared their valuable perspectives. We also extend our gratitude to the research coordinators from our "sister" study who assisted with patient recruitment, including Rebecca Greenberg, Jessica Foy, and Yuen Yu.

Funding

This research was funded by the National Center for Advancing Translational Sciences (grant R01TR004243) and the National Institute of Mental Health (5R01MH125958). The views expressed here are solely those of the authors and do not necessarily reflect the official policies of the National Institutes of Health (NIH) or the US Government.

Data Availability

The datasets presented in this article are not readily available because full datasets must remain unavailable to ensure deidentification of interview participants. Requests to access the datasets should be directed to KMKQ.

Conflicts of Interest

ES reports receiving research funding to his institution from the Ream Foundation, the International OCD Foundation, and the National Institutes of Health (NIH). He was a consultant for Brainsway and Biohaven Pharmaceuticals within the past 12 months. He owns less than US $5000 in stock in NView. He also receives book royalties from Elsevier, Wiley, Oxford, the American Psychological Association, Guildford, Springer, Routledge, and Jessica Kingsley. The remaining authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

Multimedia Appendix 1
Accuracy, validity, and trustworthiness of computer perception tools.
[DOCX File , 18 KB - mental_v13i1e79182_app1.docx]

Multimedia Appendix 2
Patient-specific relevance.
[DOCX File , 22 KB - mental_v13i1e79182_app2.docx]

Multimedia Appendix 3
Utility and implementation challenges.
[DOCX File , 17 KB - mental_v13i1e79182_app3.docx]

Multimedia Appendix 4
Regulation and governance of computer perception technologies.
[DOCX File , 25 KB - mental_v13i1e79182_app4.docx]

Multimedia Appendix 5
Data privacy and protection.
[DOCX File , 18 KB - mental_v13i1e79182_app5.docx]

Multimedia Appendix 6
Patient impacts and harms.
[DOCX File , 36 KB - mental_v13i1e79182_app6.docx]

Multimedia Appendix 7
Philosophical critiques of computer perception.
[DOCX File , 21 KB - mental_v13i1e79182_app7.docx]

References

1. Akre S, Seok D, Douglas C, Aguilera A, Carini S, Dunn J, et al. Advancing digital sensing in mental health research. NPJ Digit Med 2024 Dec 18;7(1):362 [FREE Full text] [doi: 10.1038/s41746-024-01343-x] [Medline: 39695319]
2. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017 Oct 03;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
3. Babu M, Lautman Z, Lin X, Sobota MHB, Snyder MP. Wearable devices: implications for precision medicine and the future of health care. Annu Rev Med 2024 Jan 29;75:401-415 [FREE Full text] [doi: 10.1146/annurev-med-052422-020437] [Medline: 37983384]
4. Liu JJ, Borsari B, Li Y, Liu SX, Gao Y, Xin X, et al. Digital phenotyping from wearables using AI characterizes psychiatric disorders and identifies genetic associations. medRxiv. Published online December 18, 2024 2024 Oct 18:4219 [FREE Full text] [doi: 10.1101/2024.09.23.24314219] [Medline: 39399036]
5. Orsolini L, Fiorani M, Volpe U. Digital phenotyping in bipolar disorder: which integration with clinical endophenotypes and biomarkers? Int J Mol Sci 2020 Oct 16;21(20):1-20 [FREE Full text] [doi: 10.3390/ijms21207684] [Medline: 33081393]
6. Bufano P, Laurino M, Said S, Tognetti A, Menicucci D. Digital phenotyping for monitoring mental disorders: systematic review. J Med Internet Res 2023 Dec 13;25:e46778 [FREE Full text] [doi: 10.2196/46778] [Medline: 38090800]
7. Mobbs D, Wise T, Suthana N, Guzmán N, Kriegeskorte N, Leibo JZ. Promises and challenges of human computational ethology. Neuron 2021 Jul 21;109(14):2224-2238 [FREE Full text] [doi: 10.1016/j.neuron.2021.05.021] [Medline: 34143951]
8. Torous J, Gershon A, Hays R, Onnela J, Baker JT. Digital phenotyping for the busy psychiatrist: clinical implications and relevance. Psychiatric Annals 2019 May 01;49(5):196-201. [doi: 10.3928/00485713-20190417-01]
9. D'Alfonso S, Coghlan S, Schmidt S, Mangelsdorf S. Ethical dimensions of digital phenotyping within the context of mental healthcare. Journal of Technology in Behavioral Science 2025:132-147 [FREE Full text] [doi: 10.1007/s41347-024-00423-9]

10. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019;2:88 [FREE Full text] [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]
11. Martinez-Martin N, Insel TR, Dagum P, Greely HT, Cho MK. Data mining for health: staking out the ethical territory of digital phenotyping. NPJ Digit Med 2018;1:1-5 [FREE Full text] [doi: 10.1038/s41746-018-0075-8] [Medline: 31211249]
12. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol 2017 May 08;13:23-47 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
13. Martinez-Martin N, Greely HT, Cho MK. Ethical development of digital phenotyping tools for mental health applications: Delphi study. JMIR Mhealth Uhealth 2021 Jul 28;9(7):e27343 [FREE Full text] [doi: 10.2196/27343] [Medline: 34319252]
14. Mulvenna MD, Bond R, Delaney J, Dawoodbhoy FM, Boger J, Potts C, et al. Ethical issues in democratizing digital phenotypes and machine learning in the next generation of digital health technologies. Philos Technol 2021;34(4):1945-1960 [FREE Full text] [doi: 10.1007/s13347-021-00445-8] [Medline: 33777664]
15. Shen FX, Silverman BC, Monette P, Kimble S, Rauch SL, Baker JT. An ethics checklist for digital health research in psychiatry: viewpoint. J Med Internet Res 2022 Feb 09;24(2):e31146 [FREE Full text] [doi: 10.2196/31146] [Medline: 35138261]
16. Kitson A, Marshall A, Bassett K, Zeitz K. What are the core elements of patient-centred care? A narrative review and synthesis of the literature from health policy, medicine and nursing. J Adv Nurs 2013 Jan;69(1):4-15. [doi: 10.1111/j.1365-2648.2012.06064.x] [Medline: 22709336]
17. Todres L, Galvin KT, Holloway I. The humanization of healthcare: a value framework for qualitative research. International Journal of Qualitative Studies on Health and Well-being 2009 Jul 12;4(2):68-77. [doi: 10.1080/17482620802646204]
18. Watson J. Nursing: The Philosophy and Science of Caring. Boulder, CO: Colorado University Press; 2008.
19. Shared decision-making in mental health care. Substance Abuse and Mental Health Services Administration (SAMHSA). Rockville, MD: Center for Mental Health Services, Substance Abuse and Mental Health Services Administration; 2010. URL: https://library.samhsa.gov/sites/default/files/sma09-4371.pdf [accessed 2025-05-07]
20. Insel TR. Digital phenotyping: a global tool for psychiatry. World Psychiatry 2018 Oct;17(3):276-277 [FREE Full text] [doi: 10.1002/wps.20550] [Medline: 30192103]
21. Oudin A, Maatoug R, Bourla A, Ferreri F, Bonnot O, Millet B, et al. Digital phenotyping: data-driven psychiatry to redefine mental health. J Med Internet Res 2023 Oct 04;25:e44502 [FREE Full text] [doi: 10.2196/44502] [Medline: 37792430]
22. Kostick-Quenet K, Estep J, Blumenthal-Barby JS. Ethical concerns for remote computer perception in cardiology: new stages for digital health technologies, artificial intelligence, and machine learning. Circ Cardiovasc Qual Outcomes 2024 May;17(5):e010717. [doi: 10.1161/CIRCOUTCOMES.123.010717] [Medline: 38771912]
23. Pai A, Santiago R, Glantz N, Bevier W, Barua S, Sabharwal A, et al. Multimodal digital phenotyping of diet, physical activity, and glycemia in Hispanic/Latino adults with or at risk of type 2 diabetes. NPJ Digit Med 2024 Jan 11;7(1):7 [FREE Full text] [doi: 10.1038/s41746-023-00985-7] [Medline: 38212415]
24. Provenza NR, Reddy S, Allam AK, Rajesh SV, Diab N, Reyes G, et al. Disruption of neural periodicity predicts clinical response after deep brain stimulation for obsessive-compulsive disorder. Nat Med 2024 Oct;30(10):3004-3014. [doi: 10.1038/s41591-024-03125-0] [Medline: 38997607]
25. Leaning IE, Ikani N, Savage HS, Leow A, Beckmann C, Ruhé HG, et al. From smartphone data to clinically relevant predictions: a systematic review of digital phenotyping methods in depression. Neurosci Biobehav Rev 2024 Mar;158:105541 [FREE Full text] [doi: 10.1016/j.neubiorev.2024.105541] [Medline: 38215802]
26. Onnela lab. Harvard TH Chan School of Public Health. URL: https://hsph.harvard.edu/research/onnela-lab/papers-2/ [accessed 2025-05-07]
27. Picard RW. Affective Computing. Cambridge, MA: The MIT Press; 2000.
28. Barrett LF, Mesquita B, Gendron M. Context in emotion perception. Curr Dir Psychol Sci 2011 Oct 05;20(5):286-290. [doi: 10.1177/0963721411422522]
29. Alderman JE, Palmer J, Laws E, McCradden MD, Ordish J, Ghassemi M, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. Lancet Digit Health 2025 Jan;7(1):e64-e88 [FREE Full text] [doi: 10.1016/S2589-7500(24)00224-3] [Medline: 39701919]
30. Trustworthy and responsible AI resource center - AI risks and trustworthiness. National Institute of Standards and Technology (NIST). URL: https://airc.nist.gov/airmf-resources/airmf/3-sec-characteristics/ [accessed 2025-04-16]
31. Adler DA, Stamatis CA, Meyerhoff J, Mohr DC, Wang F, Aranovich GJ, et al. Measuring algorithmic bias to analyze the reliability of AI tools that predict depression risk using smartphone sensed-behavioral data. Res Sq 2024 Apr 22:1-11 [FREE Full text] [doi: 10.21203/rs.3.rs-3044613/v1] [Medline: 38746448]
32. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: implications for clinical decision-making. PLoS Digit Health 2024 Nov;3(11):e0000651. [doi: 10.1371/journal.pdig.0000651] [Medline: 39509461]
33. Khera R, Simon MA, Ross JS. Automation bias and assistive AI: risk of harm from AI-driven clinical decision support. JAMA 2023 Dec 19;330(23):2255-2257. [doi: 10.1001/jama.2023.22557] [Medline: 38112824]

34. Walsh CG, Chaudhry B, Dua P, Goodman KW, Kaplan B, Kavuluru R, et al. Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. JAMIA Open 2020 Apr;3(1):9-15 [FREE Full text] [doi: 10.1093/jamiaopen/ooz054] [Medline: 32607482]
35. Shen FX, Baum ML, Martinez-Martin N, Miner AS, Abraham M, Brownstein CA, et al. Returning individual research results from digital phenotyping in psychiatry. Am J Bioeth 2024 Feb;24(2):69-90 [FREE Full text] [doi: 10.1080/15265161.2023.2180109] [Medline: 37155651]
36. Tabassi E. Artificial intelligence risk management framework (AI RMF 1.0). National Institute of Standards and Technology (NIST). URL: https://doi.org/10.6028/NIST.AI.100-1 [accessed 2025-05-07]
37. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union/European Union. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [accessed 2025-05-07]
38. Bahmani A. Deep data and precision health. Inside Precision Medicine 2022 Aug 01;9(4):44-46. [doi: 10.1089/ipm.09.04.12]
39. Hurley ME, Sonig A, Herrington J, Storch EA, Lázaro-Muñoz G, Blumenthal-Barby J, et al. Ethical considerations for integrating multimodal computer perception and neurotechnology. Front Hum Neurosci 2024;18:1332451 [FREE Full text] [doi: 10.3389/fnhum.2024.1332451] [Medline: 38435745]
40. Perez-Pozuelo I, Spathis D, Gifford-Moore J, Morley J, Cowls J. Digital phenotyping and sensitive health data: implications for data governance. J Am Med Inform Assoc 2021 Aug 13;28(9):2002-2008 [FREE Full text] [doi: 10.1093/jamia/ocab012] [Medline: 33647989]
41. Häuselmann A. Fit for purpose? Affective computing meets EU data protection law. International Data Privacy Law 2021 Aug;11(3):245-256. [doi: 10.1093/idpl/ipab008]
42. Lay W, Gasparini L, Siero W, Hughes EK. A rapid review of the benefits and challenges of dynamic consent. Research Ethics 2024 Sep 09;21(1):180-202. [doi: 10.1177/17470161241278064]
43. Brodkin E, Pallathra A. Missing Each Other: How To Cultivate Meaningful Connections. New York, NY: Robinson; Jan 26, 2021.
44. Stroud AM, Curtis SH, Weir IB, Stout JJ, Barry BA, Bobo WV, et al. Physician perspectives on the potential benefits and risks of applying artificial intelligence in psychiatric medicine: qualitative study. JMIR Ment Health 2025 Feb 10;12:e64414 [FREE Full text] [doi: 10.2196/64414] [Medline: 39928397]
45. Martani A, Starke G. Personal responsibility for health: the impact of digitalisation. Journal of Medical Law and Ethics 2019 Dec 31;7(3):241-258. [doi: 10.7590/221354020x15815920230933]
46. Nghiem J, Adler DA, Estrin D, Livesey C, Choudhury T. Understanding mental health clinicians' perceptions and concerns regarding using passive patient-generated health data for clinical decision-making: qualitative semistructured interview study. JMIR Form Res 2023 Aug 10;7:e47380 [FREE Full text] [doi: 10.2196/47380] [Medline: 37561561]
47. Schmidt S, D'Alfonso S. Clinician perspectives on how digital phenotyping can inform client treatment. Acta Psychol (Amst) 2023 May;235:103886 [FREE Full text] [doi: 10.1016/j.actpsy.2023.103886] [Medline: 36921359]
48. Howard M. Wearables, the marketplace and efficiency in healthcare: how will i know that you're thinking of me? Philos Technol 2021 Aug 25;34(4):1545-1568. [doi: 10.1007/s13347-021-00473-4]
49. Warraich HJ, Califf RM, Krumholz HM. The digital transformation of medicine can revitalize the patient-clinician relationship. NPJ Digit Med 2018;1:49 [FREE Full text] [doi: 10.1038/s41746-018-0060-2] [Medline: 31304328]
50. Boyatzis R. Transforming Qualitative Information: Thematic Analysis and Code Development. Thousand Oaks, CA: Sage Publications, Inc; 1998.
51. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology 2008 Jul 21;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
52. Kostick-Quenet KM, Hurley M, Herrington J, Storch EA. Rethinking ethics for an era of trusted computational tools. Psychiatric Clinics of North America 2025 Oct;1:1 (forthcoming). [doi: 10.1016/j.psc.2025.08.016]
53. Barrett LF, Adolphs R, Marsella S, Martinez AM, Pollak SD. Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. Psychol Sci Public Interest 2019 Jul;20(1):1-68 [FREE Full text] [doi: 10.1177/1529100619832930] [Medline: 31313636]
54. Birk RH, Samuel G. Digital phenotyping for mental health: reviewing the challenges of using data to monitor and predict mental health problems. Curr Psychiatry Rep 2022 Oct;24(10):523-528. [doi: 10.1007/s11920-022-01358-9] [Medline: 36001220]
55. H Birk R, Samuel G. Can digital data diagnose mental health problems? A sociological exploration of 'digital phenotyping'. Sociol Health Illn 2020 Nov;42(8):1873-1887. [doi: 10.1111/1467-9566.13175] [Medline: 32914445]
56. Chen A. A Neuroscientist explains the origins of emotions. The Verge. 2017 Apr 17. URL: https://www.theverge.com/2017/4/10/15245690/how-emotions-are-made-neuroscience-lisa-feldman-barrett [accessed 2025-04-14]
57. Le Mau T, Hoemann K, Lyons SH, Fugate JMB, Brown EN, Gendron M, et al. Professional actors demonstrate variability, not stereotypical expressions, when portraying emotional states in photographs. Nat Commun 2021 Aug 19;12(1):5037 [FREE Full text] [doi: 10.1038/s41467-021-25352-6] [Medline: 34413313]

58. Barrett LF. The theory of constructed emotion: an active inference account of interoception and categorization. Soc Cogn Affect Neurosci 2017 Jan 01;12(1):1-23 [FREE Full text] [doi: 10.1093/scan/nsw154] [Medline: 27798257]
59. Emanuel A, Eldar E. Emotions as computations. Neurosci Biobehav Rev 2023 Jan;144(1):104977-104940 [FREE Full text] [doi: 10.1016/j.neubiorev.2022.104977] [Medline: 36435390]
60. Manta C, Patrick-Lake B, Goldsack JC. Digital measures that matter to patients: a framework to guide the selection and development of digital measures of health. Digit Biomark 2020;4(3):69-77 [FREE Full text] [doi: 10.1159/000509725] [Medline: 33083687]
61. Kostick-Quenet KM, Herrington J, Storch EA. Personalized roadmaps for returning results from digital phenotyping. Am J Bioeth 2024 Feb;24(2):102-105. [doi: 10.1080/15265161.2023.2296454] [Medline: 38295237]
62. Sonig A, Deeney C, Hurley ME, Storch EA, Herrington J, Lázaro-Muñoz G, et al. What patients and caregivers want to know when consenting to the use of digital behavioral markers. NPP—Digit Psychiatry Neurosci 2024 Dec 06;2(1):1-15. [doi: 10.1038/s44277-024-00022-9]
63. Maher NA, Senders JT, Hulsbergen AFC, Lamba N, Parker M, Onnela J, et al. Passive data collection and use in healthcare: a systematic review of ethical issues. Int J Med Inform 2019 Sep;129:242-247. [doi: 10.1016/j.ijmedinf.2019.06.015] [Medline: 31445262]
64. O'Doherty KC, Christofides E, Yen J, Bentzen HB, Burke W, Hallowell N, et al. If you build it, they will come: unintended future uses of organised health data collections. BMC Med Ethics 2016 Sep 06;17(1):54 [FREE Full text] [doi: 10.1186/s12910-016-0137-x] [Medline: 27600117]
65. Segura Anaya LH, Alsadoon A, Costadopoulos N, Prasad PWC. Ethical implications of user perceptions of wearable devices. Sci Eng Ethics 2018 Feb;24(1):1-28. [doi: 10.1007/s11948-017-9872-8] [Medline: 28155094]
66. Kreitmair KV, Cho MK, Magnus DC. Consent and engagement, security, and authentic living using wearable and mobile health technology. Nat Biotechnol 2017 Jul 12;35(7):617-620. [doi: 10.1038/nbt.3887] [Medline: 28700542]
67. Lee H, Lee U. Dynamic consent for sensor-driven research. 2021 Presented at: Thirteenth International Conference on Mobile Computing and Ubiquitous Network (ICMU); November 17-19, 2021; Tokyo, Japan. [doi: 10.23919/icmu50196.2021.9638790]
68. Rake EA, van Gelder MMHJ, Grim DC, Heeren B, Engelen LJLPG, van de Belt TH. Personalized consent flow in contemporary data sharing for medical research: a viewpoint. Biomed Res Int 2017;2017:7147212 [FREE Full text] [doi: 10.1155/2017/7147212] [Medline: 28638834]
69. Basile MJ, Rubin E, Wilson ME, Polo J, Jacome SN, Brown SM, et al. Humanizing the ICU patient: a qualitative exploration of behaviors experienced by patients, caregivers, and ICU staff. Crit Care Explor 2021 Jun;3(6):e0463 [FREE Full text] [doi: 10.1097/CCE.0000000000000463] [Medline: 34151284]
70. Busch IM, Moretti F, Travaini G, Wu AW, Rimondini M. Humanization of care: key elements identified by patients, caregivers, and healthcare providers. a systematic review. Patient 2019 Oct;12(5):461-474. [doi: 10.1007/s40271-019-00370-1] [Medline: 31203515]
71. Meneses-La-Riva ME, Suyo-Vega JA, Fernández-Bedoya VH. Humanized care from the nurse-patient perspective in a hospital setting: a systematic review of experiences disclosed in Spanish and Portuguese scientific articles. Front Public Health 2021;9:737506 [FREE Full text] [doi: 10.3389/fpubh.2021.737506] [Medline: 34926369]
72. Berendt E, Tanta K. The 'heart' of things: a conceptual metaphoric analysis of heart and related body parts in Thai, Japanese and English. Intercultural Communication Studies 2011 Jan 1;20(1):7 [FREE Full text]
73. Fabian K, Fannoh J, Washington GG, Geninyan WB, Nyachienga B, Cyrus G, et al. "My heart die in me": idioms of distress and the development of a screening tool for mental suffering in Southeast Liberia. Cult Med Psychiatry 2018 Sep;42(3):684-703 [FREE Full text] [doi: 10.1007/s11013-018-9581-z] [Medline: 29728795]
74. Koydemir S, Essau C. Anxiety and anxiety disorders in young people: a cross-cultural perspective. In: Understanding Uniqueness and Diversity in Child and Adolescent Mental Health. Amsterdam, the Netherlands: Elsevier; 2018:115-134.
75. Vásquez-Londoño CA, Cubillos-Cuadrado L, Forero-Ozer A, Escobar-Espinosa P, Cubillos-López DO, Castaño-Betancur DF. Principle of hot and cold and its clinical application in Latin American and Caribbean Medicines. Adv Exp Med Biol 2021;1343:57-83. [doi: 10.1007/978-3-030-80983-6_5] [Medline: 35015277]

Abbreviations

AI: artificial intelligence
CP: computer perception
HIPAA: Health Insurance Portability and Accountability Act
REDCap: Research Electronic Data Capture

Edited by J Torous; submitted 17.Jun.2025; peer-reviewed by M Pulier; comments to author 26.Aug.2025; revised version received 20.Nov.2025; accepted 21.Nov.2025; published 05.Jan.2026.

Please cite as:
Kostick-Quenet KM, Hurley ME, Ayaz S, Herrington JD, Zampella CJ, Parish-Morris J, Tunç B, Lázaro-Muñoz G, Blumenthal-Barby J, Storch EA
Stakeholder Perspectives on Humanistic Implementation of Computer Perception in Health Care: Qualitative Study
JMIR Ment Health 2026;13:e79182
URL: https://mental.jmir.org/2026/1/e79182
doi:10.2196/79182
PMID:

Retention and Engagement in Culturally Adapted Digital Mental Health Interventions: Systematic Review of Dropout, Attrition, and Adherence in Non-Western, Educated, Industrialized, Rich, Democratic Settings

Tanya Tandon[1*], PhD; Rajashree Biswas[2*], MSc; Quentin Meteier[1,3], PhD; Karl Daher[1,4], PhD; Omar Abou Khaled[4], PhD; Björn Meyer[5], PhD; Thomas Berger[6], PhD; Rashmi Gupta[2,7], PhD; Chantal Martin Soelch[1], PhD

[1]Department of Clinical and Health Psychology, University of Fribourg, Rue P.-A.-de-Faucigny 2, Fribourg, Switzerland

[2]Cognitive and Behavioural Neuroscience Laboratory, Department of Humanities and Social Sciences, Indian Institute of Technology Bombay, Mumbai, India

[3]HumanTech Institute, HES-SO University of Applied Sciences Western Switzerland, Fribourg, Switzerland

[4]Haute École d'Ingénierie et d'Architecture de Fribourg, Fribourg, Switzerland

[5]Department of Research, Gaia AG, Hamburg, Germany

[6]Institute of Psychology, University of Bern, Bern, Switzerland

[7]Koita Centre for Digital Health, Indian Institute of Technology Bombay, Mumbai, India

[*]these authors contributed equally

Corresponding Author:
Tanya Tandon, PhD
Department of Clinical and Health Psychology, University of Fribourg, Rue P.-A.-de-Faucigny 2, Fribourg, Switzerland

Abstract

Background: Digital mental health interventions (DMHIs) offer scalable and cost-effective support for mental health but are predominantly developed in WEIRD (western, educated, industrialized, rich, democratic) contexts, raising questions about their global applicability. Dropout, attrition, and adherence rates critically influence DMHI effectiveness yet remain poorly characterized in culturally adapted formats.

Objective: This systematic review aimed to (1) synthesize evidence on dropout, attrition, and adherence in culturally adapted DMHIs delivered to non-WEIRD adult populations and (2) assess the methodological quality of the included studies.

Methods: PsycINFO, PubMed, and ScienceDirect were systematically searched for randomized controlled trials published in English between January 2014 and April 2024. Screening and data extraction followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, and methodological quality was evaluated using the Appraisal Tool for Cross-Sectional Studies tool. Extracted variables included dropout, attrition, adherence, adaptation techniques, and clinical outcomes.

Results: Twenty-three randomized controlled trials (n=4656) from diverse regions met inclusion criteria. Attrition ranged from 5.3% to 87% (median 18.4%), dropout from 0% to 66% (median 18.7%), and adherence from 26.3% to 100% (median 71%). Deep, participatory adaptations—such as language translation combined with culturally resonant content, stakeholder engagement, and iterative refinement—were consistently associated with lower dropout (<11%) and higher adherence (>75%). In contrast, surface-level adaptations (eg, translation only) showed higher dropout (up to 56%). Studies that incorporated both cultural tailoring and human support reported the most favorable engagement and clinical outcomes (eg, reductions in insomnia, depression, and anxiety). Most studies (91%) were rated as "Good" quality, although some lacked representative sampling or objective engagement metrics.

Conclusions: Comprehensive and participatory cultural adaptation is associated with engagement and effectiveness of DMHIs among non-WEIRD populations. Future research should integrate hybrid human-digital delivery models, objective engagement metrics, and larger multicenter trials to improve generalizability and scalability.

Trial Registration: PROSPERO CRD42025641863; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=641863

(JMIR Ment Health 2026;13:e80624) doi:10.2196/80624

KEYWORDS

digital mental health; cultural adaptation; non-WEIRD settings; dropout; attrition; adherence; online intervention; cultural tailoring; western, educated, industrialized, rich, democratic

Introduction

Digital mental health interventions (DMHIs) have seen explosive growth in recent years [1], offering scalable, cost-effective ways to broaden access, reduce costs, and empower users to self-manage their well-being [2,3]. These platforms—including mobile apps, video-based therapy, peer-led communities, and interactive web modules—provide flexible, on-demand support [4], with users reporting benefits such as scheduling and location flexibility, low effort, enhanced access and anonymity, greater trustworthiness with facilitators [5]. Pandemic-related demand and advances in digital access have accelerated DMHI development globally [6,7].

Alongside this expansion, longstanding debates about cultural relevance in public health have extended into the digital realm [8,9]. Most digital health research remains rooted in WEIRD (western, educated, industrialized, rich, democratic) settings [10-12], raising questions about generalizability. While some small-scale adaptations—such as a sleep-support app tailored for German refugees—have demonstrated high satisfaction and comparable adherence [13]. However, a systematic review of over 10,000 participants found no consistent efficacy advantage for culturally adapted interventions [14]. These mixed findings suggest that adaptation may boost initial uptake but does not guarantee sustained engagement or better outcomes.

A critical factor underlying these mixed results is participant retention [15]. High dropout and attrition can erode both effectiveness and cost-efficiency [16], making retention metrics essential for evaluation. Research typically focuses on three core measures: the dropout rate (discontinuation before completion [17]), the attrition rate (loss to follow-up or ceased usage [18]), and the adherence rate (completion of prescribed sessions [19]). Understanding what drives these outcomes—be it cultural fit, usability barriers, or motivational factors—is key to crafting sustainable, impactful digital interventions [20,21].

Therefore, the present systematic review aims to comprehensively examine retention and engagement outcomes in culturally adapted DMHIs implemented among non-WEIRD adult populations. Specifically, this review seeks to (1) synthesize evidence on dropout, attrition, and adherence rates across studies and (2) evaluate the methodological quality of the included trials to identify strengths, limitations, and factors associated with higher retention and adherence.

By addressing these objectives, the review intends to generate evidence-based recommendations to guide the design and implementation of culturally responsive DMHIs worldwide.

Methods

Article Search and Selection

This review was preregistered on PROSPERO (Prospective Specific Evaluation of Reviews) (CRD42025641863 [22]). The literature search took place from February 2024 and ended in July 2024. To ensure comprehensiveness, we used three search strategies: database searches and manual searches of reference lists of relevant articles.

Database Search

The review followed the guidelines of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [23]. A comprehensive search was conducted in the following electronic databases: (1) PsycINFO, (2) PubMed, and (3) ScienceDirect. The search strategy was designed to identify both quantitative and qualitative studies focusing on culturally adapted mental health interventions delivered via digital platforms (eg, e-mental health, mobile applications). Search terms were developed using combinations of keywords related to cultural adaptation (eg, "culturally appropriate," "adapted intervention") and intervention modality (eg, "digital health," "mobile app," "e-mental health"). To ensure breadth and sensitivity, the search strategy incorporated a wide range of related terms. The complete search strategy is provided in Multimedia Appendix 1 [24]. Given that previous reviews included studies up to 2014, only studies published between January 2014 and April 2024 were included. The systematic search process and the rationale for study inclusion and exclusion were documented in accordance with PRISMA standards (see Checklist 1). Two lead authors independently reviewed articles for inclusion, with disagreements resolved through discussion and consensus.

Inclusion and Exclusion Criteria

The population, intervention, control, and outcomes model served as the foundation for the creation of the inclusion criteria [25]. People from non-WEIRD societies were referred to as part of the population [12]. This systematic review focuses on DMHIs adapted for non-WEIRD populations, with clearly defined inclusion and exclusion criteria.

1. Eligible interventions must be internet-, computer-, or mobile-based to address mental health problems, including depression, anxiety, or trauma.
2. They must also be culturally adapted for the target group to align with the population's cultural context.
3. The target population includes adults aged 18 years or older from non-WEIRD cultural backgrounds that differ from the original intervention target group.
4. Only randomized controlled trials (RCTs) published in peer-reviewed English-language journals within the last 10 years are included, with no restrictions on the type of setting (eg, rural, urban, clinical, or non-clinical).

Exclusion criteria were excluded if they (1) involved interventions that lack cultural adaptation, (2) targeted individuals under 18 years, (3) were nondigital interventions, (4) were observational studies, case reports, and qualitative studies, and (5) were articles not published in English or outside the 10-year timeframe. By adhering to these criteria, the review will evaluate the impact of cultural adaptations on reducing drop-out rates and the overall effectiveness of these interventions.

Data Screening and Eligibility

After duplicates were removed using EndNote (version 20.3; Clarivate), the remaining records were screened manually using Microsoft Excel. The titles and abstracts were independently screened by two lead authors, based on pre-established inclusion and exclusion criteria. The level of agreement between the screeners was 85% across title/abstract screening, data extraction, and quality assessment stages. Discrepancies were resolved through discussion until consensus was reached.

Data Extraction

Overview

Data extraction was conducted manually using a predesigned Excel spreadsheet. The data extraction plan was developed in accordance with PRISMA guidelines and informed by recent reviews on digital health interventions among minority populations [26-28]. One author extracted the data, and another author independently cross-checked the entries for accuracy. As this is a systematic review, no imputation or sensitivity analyses were conducted. Medians and ranges were calculated only for studies that explicitly reported each outcome, and the number of contributing studies (n) is provided for each summary statistic.

Extraction of Participant Demographics

Demographic information, including participants' age, gender, and cultural background, was extracted directly from the study descriptions or participant tables. Missing or incomplete demographic data were noted in the extraction sheet.

Extraction of Recruitment Settings

Recruitment methods and settings (eg, community-based, clinical, or online) were coded from the methods section of each study. When not explicitly stated, the inferred setting was noted.

Extraction of Engagement Metrics (Dropout, Adherence, and Attrition)

Engagement data were extracted as follows: dropout was defined as noncompletion of the intervention; attrition as loss to follow-up; and adherence as the proportion of sessions completed. If data were not reported, this was recorded as "not available."

Extraction of Cultural Adaptation Strategies

Details on cultural adaptation (eg, translation, content tailoring, stakeholder involvement, and iterative feedback) were extracted from intervention descriptions. Adaptations were coded as surface-level or deep-level.

Extraction of Clinical Outcomes

Primary and secondary clinical outcomes (eg, depression, anxiety, insomnia) and their corresponding measurement tools were extracted and coded for direction of effect (improvement, no change, or worsening).

Quality Assessment

Two authors independently conducted the quality assessment of all included quantitative studies using the Appraisal Tool for Cross-Sectional Studies [29]. Disagreements were resolved by discussion. Each item was rated as "yes," "no," or "do not know," with scores assigned according to conventions used in previous reviews [30,31]: yes or not applicable (N/A)=1 point; no or do not know=0 points. Total scores ranged from 0 to 20, with studies rated as good (≥15), fair (10-14), or poor (<10).

Results

Study Selection

A total of 184,047 records were identified through the database search. After removing duplicates (n=180,371), 3676 articles remained for title and abstract screening. Of these, 3641 were excluded based on the predefined inclusion and exclusion criteria. The remaining 35 articles were assessed for eligibility by the authors. Eleven studies were excluded at this stage because they were study protocols or review articles, and one study met all inclusion criteria but was excluded from the final review due to inaccessibility; attempts to obtain the full text through institutional subscriptions and direct author contact were unsuccessful. Ultimately, 23 articles were included in the final review. The study selection process is summarized in Figure 1, and detailed study descriptions are available in Multimedia Appendix 2 and Table 1.

Figure 1. PRISMA flowchart. RCT: randomized controlled trial.

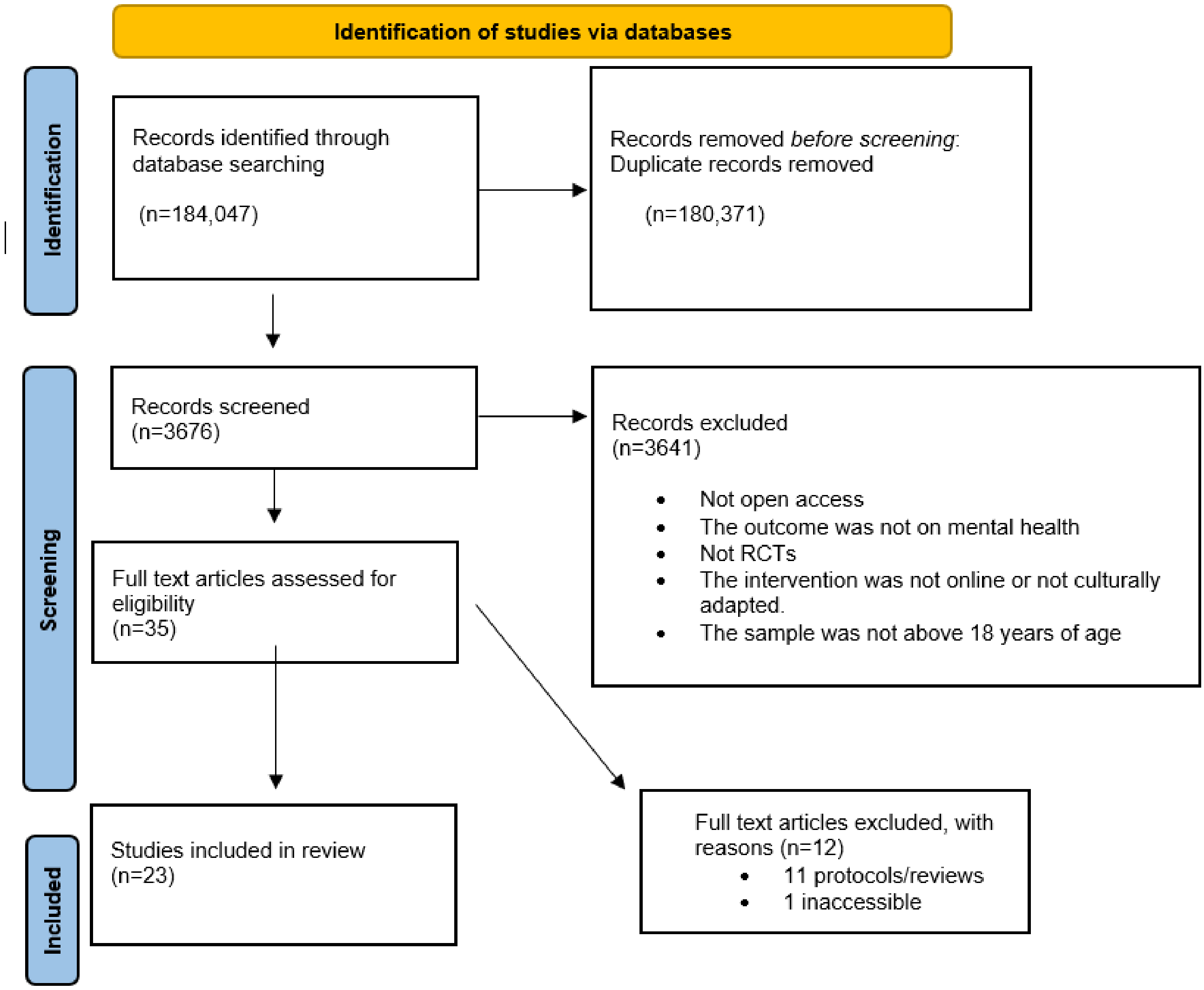


Table . Study description of the selected studies (n=23).

Author and year	Country of origin	Adapted country	Recruitment settings	Age (y), mean (SD)	Demographics	Intervention type	Platform used	Adaptation framework used	Dropout rate (%)	Attrition rate (%)	Adherence rate (%)	Primary outcome measure
Zhang et al [32] (2023)	China	China	Not reported	49.67 (14.49)	Insomnia (chronic insomnia disorder), Chinese; 74.4% female	DCBT-I[a] app	Smartphone-based app	Not reported	11	11	94	Insomnia (ISI[b])
Spanhel et al [13] (2022)	Germany	Germany	Online	26.8 (4.4)	International students in Germany (92.6% with insomnia); 49.4% female	StudiCare Sleep-e based on CBT[c]	Web-based intervention on Minddistrict platform	Adaptation included content (eg, removal of sleep restriction), duration (shortened from 6 to 3 modules), language (translation into English), and the use of students as case examples.	56	46	44	Insomnia (ISI)
Zeng et al [33] (2020)	China	China	Outpatient clinics; online	28 (5.8)	HIV seropositive individuals with depressive symptoms; 5.33% female.	WeChat-based mHealth[d] intervention	WeChat-based (app-based)	Not reported	50	8	100	Depression (CES-D)[e]
Guo et al [34] (2022)	China	China	Outpatient clinics; online	28.3 (5.85)	HIV seropositive individuals with depressive symptoms; 7.65% female.	Run4Love based on CBSM[f]	WeChat with multimedia materials, automated tracking, and phone check-ins.	CBSM adapted in Chinese context	41	41	50	Depression (CES-D)[e]

Author and year	Country of origin	Adapted country	Recruitment settings	Age (y), mean (SD)	Demographics	Intervention type	Platform used	Adaptation framework used	Dropout rate (%)	Attrition rate (%)	Adherence rate (%)	Primary outcome measure
Campbell et al [35] (2023)	United States	United States	Outpatient clinics	38. 6 (10.3)	American Indian and Alaska Native in the United States; 45.3% female, 1.9% transgender.	TES-NAV[g]	Smartphone app or clinic tablets		49	31	74	Abstinence from heavy drinking or drug use (urine screen and self-report)

Author and year	Country of origin	Adapted country	Recruitment settings	Age (y), mean (SD)	Demographics	Intervention type	Platform used	Adaptation framework used	Dropout rate (%)	Attrition rate (%)	Adherence rate (%)	Primary outcome measure
								Integrated multi-framework adaptation: (1) Ecological Validity Model (Bernal et al [36], 2009) to align language, persons, metaphors, content, concepts, methods, goals, and context; (2) Barrera et al's [37] (2013) 5-step systematic cycle (information gathering → preliminary design → pilot test → refinement → final trial); and (3) Wingood and DiClemente's [38] (2008) cultural-tailoring principles. All steps were iteratively co-designed with four native clinicians/psychologists, individuals				

Author and year	Country of origin	Adapted country	Recruitment settings	Age (y), mean (SD)	Demographics	Intervention type	Platform used	Adaptation framework used	Dropout rate (%)	Attrition rate (%)	Adherence rate (%)	Primary outcome measure
								with lived experience, and community reviewers.				
Lindegaard et al [39] (2019)	Sweden	Sweden	Not reported	33.86 (8.2)	Depressive Kurdish people of Sweden; 46% female	ICBT[h]	Secure online platform: Iterapi	No formal framework cited	28	44	52	Depression (BDI-II)[i]
Silva et al [40] (2020)	United States	United States (Spanish speaking population)	Not reported	42.7 (11.6)	Native Spanish speaking individuals; DSM[j] IV abuse and substance dependence; 32.6% female	CBT4CBT[k]	Web based	Cultural constructs by Anez et al [41,42] (2005, 2008)	12	5	88	Change in SUD[l] (ASI[m])
Yeung et al [43] (2016)	United States	United States (Chinese American immigrants)	Online	50 (14.5)	Monolingual Chinese Americans with depression, 63% female	T-CSCT[n]	Polycom VSX3000 systems were used for videoconferencing. Later, switched to Skype	Culturally sensitive psychiatric consultation using the Engagement Interview Protocol (EIP).	Not stated	Not stated	Not stated	Depression (HDRS17)[o]
Sarfraz et al [44] (2023)	Pakistan	Pakistan	Online	22.90 (3.57)	Undergraduate and postgraduate university students, 69% female	MTC[p]	Zoom and email	Medical Research Council (MRC) guidelines for complex interventions; Heuristic framework for cultural adaptation	28	28	48	Psychological distress (CORE-OM)[q]; psychological well-being (PWB-S)[r]; dispositional mindfulness (FFMQ[s])

Author and year	Country of origin	Adapted country	Recruitment settings	Age (y), mean (SD)	Demographics	Intervention type	Platform used	Adaptation framework used	Dropout rate (%)	Attrition rate (%)	Adherence rate (%)	Primary outcome measure
Zemestani and Fazeli Nikoo [45] (2020)	Iran	Iran	Not reported	29.59 (3.59)	Pregnant women (1 - 6 wk of gestational age)	MBCT[t]	In-person group sessions + audio for home practice (offline)	No formal framework cited	13	23	87	Depression (BDI-II); Anxiety (BAI)[u]; emotional regulation (ERQ)[v]; well-being (SP-WB)[w]
Spruill et al [46] (2021)	United States	United States	Outpatient clinics	43.3 (11.3)	Hispanic ethnicity; 67% primary Spanish speaker; 71% female	Project UP-LIFT[x], adapted from MBCT	Telephone	No formal framework cited—adaptations informed by qualitative research and best practices (eg, focus groups, simplification, cultural tailoring);	14	7	75	Depression (PHQ-9)[y]
Zhang et al [47] (2023)	China	China	Outpatient clinics; online	30.29 (4.29)	Pregnant women in China	GSH-MBI[z]	WeChat mini-program	No formal framework cited—adaptations relied on culturally tailored content delivered through WeChat.	19	16	81	Depression (EPDS)[aa]; Anxiety (GAD-7)[ab]
Benjet et al (2023) [48]	Mexico and Colombia	Mexico and Colombia	Online	21.4 (3.2)	University students; 1038 women (78.7%); 725 participants (55.0%) came from Mexico	i-CBT[ac]	Web based	Iterative user-centered model	Not reported	32	Not reported	Anxiety (GAD-7) and depression (PHQ-9) scores

Author and year	Country of origin	Adapted country	Recruitment settings	Age (y), mean (SD)	Demographics	Intervention type	Platform used	Adaptation framework used	Dropout rate (%)	Attrition rate (%)	Adherence rate (%)	Primary outcome measure
Vaca et al [49] (2023)	United States	US Latino adults	Not reported	36.2 (11.2)	433 (51.5%) were male, 407 (48.5%) were female and 83% of them were from Puerto Rico	AB-CASI[ad]	Computer tablets (iPad 4th Generation; Apple Inc)	Not explicitly named	Not reported	24	Not reported	Alcohol Use Disorders (AUDIT)[ae]
Zhou et al [50] (2022)	United States	United States (specific adaptation for Black women in the United States)	Not reported	59.5 (8)	American Black women	SHUTi-BWHS[af] based on CBT-I[ag]	Web-based	Stakeholder-informed, iterative cultural adaptation process (not explicitly a formal framework, but uses participatory design principles)	22	16	78	Insomnia (ISI)
Javier et al [51] (2025)	United States.	Filipino families	Not reported	42 (5.6)	Filipino; parents: 81.7% females, 16.3% males	Incredible Years School Age Basic and Advance Programs	Web based	Language, persons, metaphors, content, concepts, goals, methods, and context based on Bernal et al, [52] (1995) framework including language, persons, metaphors, content, concepts, goals, methods, context	18	18	Not explicitly mentioned	Parenting practices (PPI)[ah]; Parenting stress (PSI)[ai]; Child's behavior (CBCL)[aj]; Child-reported anxiety and depression symptoms (SCARED[ak] and CDI[al])

Author and year	Country of origin	Adapted country	Recruitment settings	Age (y), mean (SD)	Demographics	Intervention type	Platform used	Adaptation framework used	Dropout rate (%)	Attrition rate (%)	Adherence rate (%)	Primary outcome measure
Owen et al [53] (2022)	United States	African Americans in the United States	Outpatient clinics; online	65.9	African Americans; 14 females and 3 males	CBT	In-person and online group formats	Agricultural Coping Model using amalgam of norms from West Africa, cultural traditions and practices from European/American society, and experiences of historical and contemporary racism in the United States	Not reported	18	Not reported	Cognitive function (MoCA)[am]
Lindegaard et al [54] (2021)	Sweden	Sweden (for Arabic-speaking immigrants and refugees)	Online	37.5 (11.4)	Arabic-speaking population, 25 females and 34 males	ICBT	Web-based with asynchronous therapist messaging and feedback	Iterative adaptation and tailoring process (focus groups + pilot feedback)	39	Not reported	Not reported	Depression (PHQ-9)
Yamaguchi et al [55] (2019)	Japan	Japan	Community	20.25 (1.31)	University students, 26 females and 70 males	FSC[an]; IBSS[ao]	In-person (initial session)+ email follow-up	Not reported	Not reported	28	Not reported	Reported behavior, and the other on intended behavior (RIBS-J)[ap]; Mental Illness and Disorder Understanding (MIDUS)[aq]
Ellis et al [56] (2022)	United States	Egypt	Not reported	Range 20 - 54 (28)	Arabic speaking population, 62 females and 25 males	CBT-based PTSD[au] for online coaching in Arabic	Web based	Bernal et al (1995) [52] framework	Not reported	13	Not reported	PTSD (PCL-5)[ar]

Author and year	Country of origin	Adapted country	Recruitment settings	Age (y), mean (SD)	Demographics	Intervention type	Platform used	Adaptation framework used	Dropout rate (%)	Attrition rate (%)	Adherence rate (%)	Primary outcome measure
Sun et al [57] (2022)	United States	China	Online	22.21 (2.67)	University students, 73.7% females	Mindfulness-based mHealth	Web-based via WeChat (mini-program) and Zoom	Informed by focus group input; rapid iterative tailoring (no formal framework named)	9	13	Not reported	Anxiety (GAD-7); depression (PHQ-9)
Jacobs et al [58] (2016)	United States	Ecuador	Community	≥18	Students	Familias Unidas is a parent-centered intervention	In-person sessions; group and family-based, Audio Computer-Assisted Self-Interviewing	Barrera et al's [59] (2017) surface-structure adaptation model—constructs vetted against Ecuadorian family norms/laws; parent review for linguistic clarity; minor wording and local prevalence data updates; original Hispanic-acted skill videos retained, with no deep-structure changes required	Not reported	Not reported	Not reported	Drug use (self-reported by adolescents); Adolescent sexual behavior, drug use, and violence (ACASI)[as]

Author and year	Country of origin	Adapted country	Recruitment settings	Age (y), mean (SD)	Demographics	Intervention type	Platform used	Adaptation framework used	Dropout rate (%)	Attrition rate (%)	Adherence rate (%)	Primary outcome measure
Barrera et al [60] (2015)	United States	Spain	Online	30.19 (5.57)	Pregnant women, majority resided in Chile, Spain, Argentina, Mexico, Colombia, and the United States. Most were Spanish speaking (82.9%) of Latino/Hispanic ethnic identity (71.3%), and identified their racial background as Caucasian/European (53.2%) or Mestizo (31.8%)	Mothers and Babies Internet Course/Curso Internet de Mamás y Bebés (e-MB) based on CBT approach	Web-based accessed via email login links	Iterative user-feedback model (usability testing, linguistic translation, visual editing)				

[a]DCBT-I: digital cognitive behavioral therapy for insomnia.

[b]ISI: Insomnia Severity Index.

[c]CBT: cognitive behavioral therapy.

[d]mHealth: mobile health.

[e]CES-D: Center for Epidemiological Studies Depression Scale.

[f]CBSM: cognitive behavioral stress management.

[g]TES-NAV: therapeutic education system-native version.

[h]ICBT: inference-based cognitive behavioral therapy.

[i]BDI-II: Beck Depression Inventory-II.

[j]*DSM*: *Diagnostic and Statistical Manual*.

[k]CBT4CBT: Web-based cognitive behavioral therapy program.

[l]SUD: substance use disorder.

[m]ASI: Addiction Severity Index.

[n]T-CSCT: telepsychiatry-based culturally sensitive collaborative treatment.

[o]HDRS17: Hamilton Depression Rating Scale.

[p]MTC: online mindfulness training course.

[q]CORE-OM: clinical outcomes routine evaluation-outcome measure.

[r]PWB-S: Ryff's psychological well-being scale.

[s]FFMQ: Five-Facet Mindfulness Questionnaires.

[t]MBCT: mindfulness-based cognitive therapy.

[u]BAI: Beck Anxiety Inventory.
[v]ERQ: Emotion Regulation Questionnaire.
[w]SPWB: scales of psychological well-being.
[x]UPLIFT: using practice and learning to increase favorable thoughts.
[y]PHQ-9: Patient Health Questionnaire.
[z]GSH-MBI: digital guided self-help mindfulness-based intervention.
[aa]EPDS: Edinburgh Postnatal Depression Scale.
[ab]GAD-7: Generalized Anxiety Disorder.
[ac]i-CBT: Internet-delivered cognitive behavioral therapy.
[ad]AB-CASI: automated bilingual computerized alcohol screening and intervention.
[ae]AUDIT: Alcohol Use Disorders Identification Test.
[af]SHUTi-BWHS: tailored version of automated internet-delivered treatment called Sleep Healthy Using the Internet for Black women.
[ag]CBT-I: cognitive behavioral therapy for insomnia.
[ah]PPI: Parenting Practices Interview.
[ai]PSI: Parenting Stress Index.
[aj]CBCL: child behavior checklist.
[ak]SCARED: parent's screening for child anxiety-related disorders.
[al]CDI 2: parent and child report Children’s Depression Inventory 2.
[am]MoCA: Montreal cognitive assessment.
[an]FSC: filmed social contact.
[ao]IBSS: internet-based self-study.
[ap]RIBS-J: reported and intended behavior scale – Japanese version.
[aq]MIDUS: mental illness and disorder understanding scale.
[ar]PCL-5: post-traumatic stress disorder checklist.
[as]ACASI: adolescent sexual behavior, drug use, and violence.
[at]MDE: Major Depressive Episode Screener—current/lifetime version.
[au]PTSD: post-traumatic stress disorder.

Participant Demographics

Participants (n=4656; see Multimedia Appendix 2) represented diverse cultural and demographic backgrounds. Several studies focused on specific subpopulations like pregnant women who were targeted in two studies [47,60,61], while young adults aged 18 to 30 years were the focus of 39% (8/23) [33,34,44,48,55,57,61,62]. Chinese participants were the most frequently represented cultural group, included in 26% (6/23) of studies [32-34,43,47,57]. University students from various countries were the focus in 21% (5/23) of studies [44,48,55,57,62]. Studies involving Middle Eastern or Arabic-speaking populations accounted for 17% (4/23) [13,54,56,61]. A significant proportion of studies examined Hispanic/Latinx participants (7/23, 30%) [40,46,48,49,58,60,62] and only two studies (2/23, 9%) included Black American participants [50,53], and one study (1/23, 4%) evaluated a DMHI among Indigenous communities [35], highlighting ongoing underrepresentation of these groups in culturally adapted digital mental health research, and one study was conducted on Japanese students [55].

Recruitment Settings

Among the 23 included studies, most investigated DMHIs among participants residing in urban settings (8/23, 35%), typically located near metropolitan areas [32,35,43,46,49,50,55,58]. Most of the studies (14/23, 61%) often relied on community-based recruitment methods such as advertisements, mailing lists, and outreach through community centers [32,35,43,44,46,48,50,53-55,57,58,61,62]. Internet-based recruitment was the second most common strategy, used in 22% (5/23) of studies, primarily through platforms such as social media [44,54,55,57,60]. Seven studies (7/23, 30%) recruited participants directly from outpatient clinical settings located in urban areas [32-35,43,46,53]. A notable proportion of studies (7/23, 30%) recruited through universities and online platforms—to enhance sample diversity and reach [32,34,47,48,53,54,62]. For six studies (6/23, 26%), recruitment settings were not clearly reported, although some recruitment strategies (eg, convenience or snowball sampling) were described [45,47,50,51,56,61].

Engagement Metrics: Dropout, Adherence, and Attrition Rates

Across the 23 included studies, participant engagement varied substantially. Attrition rates—defined as loss to follow-up—ranged from 5.3% to 87%, with a median attrition rate of approximately 18.4% among studies reporting this outcome. While some studies demonstrated relatively low attrition (eg, <15%) [46,47,56,57,61], some reported notably high rates (>35%) [44,60,62], and five studies did not state attrition rates [33,34,43,54,58], limiting comprehensive comparison. Attrition rates were reported in 61% (14/23) of studies. Dropout rates, reflecting noncompletion of the intervention, also varied widely, from 0% to 66%, with a median dropout rate of 18.7%. Dropout rates were reported in 17 studies (17/23). Adherence rates, or the proportion of sessions or modules completed by participants, ranged from 26.3% to 100%, with a median adherence rate of approximately 71% in studies that reported these data. Adherence rates were reported in 61% (14/23) of studies. High adherence was reported in programs such as internet-delivered cognitive behavioral therapy (Sleep Healthy Using the Internet for Black women) [50], where over 60% of participants completed all modules. However, 39% (9/23) of studies did not report adherence rates [43,48,51,53-58]

Cultural Adaptation Strategies

Across the included studies, a wide range of cultural adaptation strategies were employed to enhance the relevance and effectiveness of DMHIs for diverse populations. One of the forms of cultural adaptation used in the studies was language translation, implemented in 57% (13/23) of studies to improve linguistic accessibility [32,33,44,46-48,50,51,54,56-58,60]. The other forms of cultural adaptation included content and imagery adaptations that were mainly used in 70% (16/23) of studies to align with cultural norms, such as visuals and metaphors tailored for specific populations [33,35,44,46-48,50,51,53,54,56-58,60-62]. In 70% (16/23) of studies, cultural values and local practices were integrated into the intervention design, including the incorporation of traditional healing methods for Indigenous groups [33,35,44,46-48,50,51,53,54,56-58,60-62]. Stakeholder involvement—including collaboration with cultural experts, local communities, and leaders—was reported in 48% (11/23) of studies [33,35,44,46-48,50,51,53,54,56-58,60-62]. Iterative feedback and refinement processes—using focus groups, cognitive interviews, and pilot trials—were used in 43% (10/23) of studies to adjust the interventions based on user responses [33,35,44,46-48,50,51,53,54,56-58,60-62]. Only three studies (3/23, 13%) employed the Ecological Validity Framework (EVF), guiding systematic adaptation across multiple cultural dimensions [51,56,61]. Similarly, surface- and deep-structure adaptations—which modify both observable aspects like language and deeper cultural constructs—were applied in 9% (2/23) of studies [44,58]. Technology adaptation to locally preferred platforms (eg, WeChat in China) was reported in 35% (8/23) of studies [32,33,47,56-58,60,61]. However, only one study (1/23; 4%) included cultural competency training for providers to ensure culturally sensitive delivery [53].

Clinical Outcomes

The studies included in this systematic review reported various clinical outcomes, focusing on improvements in mental health symptoms, quality of life, and other relevant measures. Most commonly, the studies targeted insomnia and sleep-related issues as primary clinical outcomes (4/23, 17%) [32,50,54,62], followed by depression (12/23, 52%) [32-34,46-49,54,57,60-62] and anxiety (8/23, 35%) [32,45,47-49,54,56,61,62]. Other notable outcomes included significant reductions in pregnancy-related anxiety among pregnant women in China using a digital guided self-help mindfulness-based intervention [47] an automated bilingual digital health tool in the United States significantly reduced binge drinking episodes [49].

Methodological Quality Assessment

Overall, most studies (21/23, 91%) were of high methodological quality. Overall, most studies demonstrated clear research aims and employed study designs that were appropriate and well justified in relation to their objectives. The target populations were clearly defined across all studies. Statistical methods were generally well described. Additionally, the key findings of the studies were usually presented clearly, with discussions and conclusions that were largely justified. Most studies also acknowledged their limitations, enhancing transparency. However, improvements are needed in two studies [58,62] by ensuring representative sampling, justifying sample sizes, addressing nonresponse bias, and transparently reporting dropout data.

Discussion

Principal Findings

This systematic review synthesized findings from 23 RCTs examining dropout, attrition, and adherence in culturally adapted DMHIs among non-WEIRD adult populations. Participant engagement varied widely, with median dropout and attrition rates around 18% and mean adherence of 71%. Interventions using deep, participatory forms of cultural adaptation—combining translation with locally meaningful content, stakeholder involvement, and iterative refinement—showed the highest adherence (often >75%) and lowest dropout (typically <11%). In contrast, interventions limited to surface-level adaptations such as language translation alone frequently exhibited higher dropout (up to 56%) and lower adherence.

Patterns in Engagement

Dropout rates ranged from 6% to 87% and appeared to vary depending on adaptation depth [47,60]. Studies integrating multiple culturally grounded elements (eg, language, imagery, values, and delivery context) reported greater retention and engagement. For instance, Zhang et al [32] integrated culturally specific sleep concepts into a CBT-I intervention, achieving a dropout rate of only 6.09%, while Silva et al [40] used culturally resonant telenovela-style content and reported dropout at 8.4%. These findings indicate that culturally resonant content may be linked to greater trust, relevance, and sustained participation.

Impact of Adaptation Depth

By contrast, interventions that employed surface-level adaptations—such as translation without deeper contextual integration—or lacked explicit cultural adaptation tended to show higher dropout. Spanhel et al [62], for example, provided a non-adapted English CBT intervention to a diverse population and observed a dropout rate of 56%. Similarly, Zeng et al [33] and Guo et al [34] implemented basic linguistic and platform-level adaptations but reported dropout rates of 49.2% and 41%, respectively. These findings suggest that surface-level efforts were typically associated with lower sustained engagement in culturally diverse populations.

Participatory Design and Implementation

The role of participatory design processes emerged as another important determinant of adherence. Studies like Zhou et al [50] and Lindegaard et al [39] used stakeholder input such as including collaboration with cultural experts, local communities, and leaders and iterative design such as using focus groups, cognitive interviews, and pilot trials, which corresponded to relatively low dropout rates (10.5% and 28%, respectively). However, participatory adaptation alone did not guarantee low attrition, as seen in Barrera et al [60], where despite iterative feedback mechanisms, dropout peaked at 86.97%, possibly due to high geographic and contextual diversity or technological barriers. This highlights the need to complement participatory design with context-sensitive implementation strategies.

Engagement and Clinical Outcomes

An overall pattern emerged in which studies with lower dropout rates were often observed alongside stronger clinical outcomes. For instance, Zhang et al [32], Silva et al [40], and Sun et al [57] demonstrated both high retention and significant reductions in insomnia, depression, or anxiety. However, some studies with moderate to high dropout (eg, Refs [44,48]) still reported clinical improvements among completers, indicating that while adaptation enhances effectiveness, it may not be sufficient to retain all users without additional strategies to address barriers to access and sustained use.

Despite engagement data from 23 RCTs, a meta-analysis was not possible due to heterogeneity in interventions, populations, outcomes, and definitions of engagement, as well as limited extractable data (≈60%) and small moderator subgroups. We therefore share the descriptive synthesis, and future work using standardized metrics may allow meta-regression.

Strengths and Limitations of Current Evidence

The reviewed studies highlight several strengths of culturally adapted interventions in supporting engagement and clinical outcomes. Many adapted programs demonstrated higher completion and retention rates, such as tailored versions of Sleep Healthy Using the Internet for Black women [50] and Project Using Practice and Learning to Increase Favorable Thoughts for Hispanic adults [46]. Several interventions also reported improved clinical outcomes, including reduced substance use, greater abstinence, and enhanced sleep or mood symptoms. These positive patterns were most often observed in studies incorporating language congruence and cultural values such as communal participation and sensitivity to race-based stressors.

Despite these successes, several limitations emerged. Many pilot trials had small samples, limiting generalizability and highlighting the need for larger, multi-center validation [32,39,46,53,54]. Perceived cultural relevance also varied within target groups; for instance, in [53], although 84% of participants found the adaptation relevant, some viewed it as "overdone," reflecting within-group diversity and the importance of facilitator racial matching [53]. Some culturally adapted DMHIs faced challenges in sustaining engagement, such as the online mindfulness course for Pakistani students with high attrition [44] and the Egyptian post-traumatic stress disorder intervention whose participants desired more "human" interaction [56]. A further limitation is that a quantitative meta-analysis or meta-regression was not performed. Considerable heterogeneity in study design, intervention type, and outcome measures—along with inconsistent definitions of adherence and dropout and limited extractable numerical data—made statistical aggregation inappropriate. Subgroup counts were also too small for stable moderator modeling. Moreover, standardized mean differences or confidence intervals for clinical outcomes could not be reported, as most studies used heterogeneous measures and lacked sufficient statistical detail. Consequently, clinical outcomes were synthesized narratively to reflect overall improvement of trends across interventions. Future research should standardize engagement metrics and reporting to enable robust meta-analytic and meta-regression approaches that can better quantify determinants of adherence and attrition. Finally, many studies relied solely on self-reported outcomes and unblinded data collection, increasing the risk of bias [53,56]

Conclusion

Drawing from these implications, several key recommendations emerge for future research and practice. First, it is essential to prioritize comprehensive cultural adaptation, moving beyond superficial changes to genuinely embed content and delivery methods within the target culture's values and sociocultural realities [53]. This includes ensuring language congruence [46] and actively involving community members and cultural experts in the design process to ensure adaptations are relevant and address within-group heterogeneity [53]. Second, to support [22] engagement and retention, hybrid models integrating human support should be considered, as noted by users of a culturally adapted web-based post-traumatic stress disorder intervention for Egyptians who desired more "human" interaction and personalization [56]. Proactive monitoring of engagement metrics is also vital to enable timely re-engagement strategies [33]. Third, given that the perceived cultural relevance can be influenced by the race of the intervention facilitator [53], comprehensive cultural competence and implicit bias training for facilitators are recommended to build trust and address potential microaggressions [53].

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This study is funded by Swiss National Science Foundation: Trilateral Call with India (IZINZ1_209481) and Indian Council of Social Science Research (grant no.: RD/0123-ICSSR00-003).

Data Availability

The datasets that will be used and/or analyzed during the current study will be made available from the corresponding author on reasonable request.

Authors' Contributions

TT and RB screened the studies. TT and RB wrote the first draft of the manuscript. TT and RB jointly revised the manuscript based on the comments given by TB, CMS, QM, KD, BM, and RG. All authors contributed to the article and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete search strategy used for PsycINFO, PubMed, and ScienceDirect databases.
[DOCX File, 16 KB - mental_v13i1e80624_app1.docx]

Multimedia Appendix 2
Detailed study descriptions (Table 1) of included randomized controlled trials.
[XLSX File, 65 KB - mental_v13i1e80624_app2.xlsx]

Checklist 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File, 32 KB - mental_v13i1e80624_app3.docx]

References

1. Graham AK, Lattie EG, Powell BJ, et al. Implementation strategies for digital mental health interventions in health care settings. Am Psychol 2020 Nov;75(8):1080-1092. [doi: 10.1037/amp0000686] [Medline: 33252946]
2. Koh JH, Chong LCY, Koh GCH, Tyagi S. Telemedical interventions for chronic obstructive pulmonary disease management: umbrella review. J Med Internet Res 2023 Feb 16;25:e33185. [doi: 10.2196/33185] [Medline: 36795479]
3. Weatherly S, McKenna T, Wahba S, et al. Effectiveness of digital health interventions (DHI) in chronic pain management: a scoping review of current evidence and emerging trends. Cureus 2024 Oct;16(10):e72562. [doi: 10.7759/cureus.72562] [Medline: 39610577]
4. Early J, Gonzalez C, Gordon-Dseagu V, Robles-Calderon L. Use of mobile Health (mHealth) technologies and interventions among community health workers globally: a scoping review. Health Promot Pract 2019 Nov;20(6):805-817. [doi: 10.1177/1524839919855391] [Medline: 31179777]
5. Wallin EEK, Mattsson S, Olsson EMG. The preference for internet-based psychological interventions by individuals without past or current use of mental health treatment delivered online: a survey study with mixed-methods analysis. JMIR Ment Health 2016 Jun 14;3(2):e25. [doi: 10.2196/mental.5324] [Medline: 27302200]
6. Wang Q, Su M, Zhang M, Li R. Integrating digital technologies and public health to fight COVID-19 pandemic: key technologies, applications, challenges and outlook of digital healthcare. IJERPH 2021;18(11):6053. [doi: 10.3390/ijerph18116053]
7. Abernethy A, Adams L, Barrett M, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022;2022:10. [doi: 10.31478/202206e] [Medline: 36177208]
8. Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z. Digital health is a cultural transformation of traditional healthcare. mHealth 2017;3:38. [doi: 10.21037/mhealth.2017.08.07] [Medline: 29184890]
9. Naderbagi A, Loblay V, Zahed IUM, et al. Cultural and contextual adaptation of digital health interventions: narrative review. J Med Internet Res 2024 Jul 9;26:e55130. [doi: 10.2196/55130] [Medline: 38980719]
10. Tandon T, Piccolo M, Ledermann K, Gupta R, Morina N, Martin-Soelch C. Relationship between behavioral and mood responses to monetary rewards in a sample of Indian students with and without reported pain. Sci Rep 2022 Nov 24;12(1):20242. [doi: 10.1038/s41598-022-24821-2] [Medline: 36424426]
11. Tandon T, Piccolo M, Ledermann K, et al. Mental health markers and protective factors in students with symptoms of physical pain across WEIRD and non-WEIRD samples - a network analysis. BMC Psychiatry 2024 Apr 24;24(1):318. [doi: 10.1186/s12888-024-05767-3] [Medline: 38658915]
12. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci 2010 Jun;33(2-3):61-83. [doi: 10.1017/S0140525X0999152X] [Medline: 20550733]
13. Spanhel K, Hovestadt E, Lehr D, et al. Engaging refugees with a culturally adapted digital intervention to improve sleep: a randomized controlled pilot trial. Front Psychiatry 2022;13:832196. [doi: 10.3389/fpsyt.2022.832196] [Medline: 35280163]
14. Spanhel K, Balci S, Feldhahn F, Bengel J, Baumeister H, Sander LB. Cultural adaptation of internet- and mobile-based interventions for mental disorders: a systematic review. NPJ Digit Med 2021 Aug 25;4(1):128. [doi: 10.1038/s41746-021-00498-1] [Medline: 34433875]

15. Borghouts J, Eikey E, Mark G, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
16. 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 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
17. Postel MG, de Haan HA, ter Huurne ED, Becker ES, de Jong CAJ. Effectiveness of a web-based intervention for problem drinkers and reasons for dropout: randomized controlled trial. J Med Internet Res 2010 Dec 16;12(4):e68. [doi: 10.2196/jmir.1642] [Medline: 21163776]
18. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11. [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
19. Wangberg SC, Bergmo TS, Johnsen JAK. Adherence in Internet-based interventions. Patient Prefer Adherence 2008 Feb 2;2:57-65. [Medline: 19920945]
20. Skea ZC, Newlands R, Gillies K. Exploring non-retention in clinical trials: a meta-ethnographic synthesis of studies reporting participant reasons for drop out. BMJ Open 2019 Jun 3;9(6):e021959. [doi: 10.1136/bmjopen-2018-021959] [Medline: 31164359]
21. Amagai S, Pila S, Kaat AJ, Nowinski CJ, Gershon RC. Challenges in participant engagement and retention using mobile health apps: literature review. J Med Internet Res 2022 Apr 26;24(4):e35120. [doi: 10.2196/35120] [Medline: 35471414]
22. Retention and engagement in culturally adapted digital mental health interventions: a systematic review of dropout, attrition, and adherence in non-WEIRD settings. National Institute for Health and Care Research. URL: https://www.crd.york.ac.uk/PROSPERO/view/CRD42025641863 [accessed 2026-01-10]
23. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097. [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
24. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ ;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
25. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak 2007 Jun 15;7:16. [doi: 10.1186/1472-6947-7-16] [Medline: 17573961]
26. Armaou M. Research trends in the study of acceptability of digital mental health-related interventions: a bibliometric and network visualisation analysis. Soc Sci (Basel) 2024;13(2):114. [doi: 10.3390/socsci13020114]
27. Harper Shehadeh M, Heim E, Chowdhary N, Maercker A, Albanese E. Cultural adaptation of minimally guided interventions for common mental disorders: a systematic review and meta-analysis. JMIR Ment Health 2016 Sep 26;3(3):e44. [doi: 10.2196/mental.5776] [Medline: 27670598]
28. Ramos G, Chavira DA. Use of technology to provide mental health care for racial and ethnic minorities: evidence, promise, and challenges. Cogn Behav Pract 2022 Feb;29(1):15-40. [doi: 10.1016/j.cbpra.2019.10.004]
29. Downes MJ, Brennan ML, Williams HC, Dean RS. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ Open 2016 Dec 8;6(12):e011458. [doi: 10.1136/bmjopen-2016-011458] [Medline: 27932337]
30. Moor L, Anderson JR. A systematic literature review of the relationship between dark personality traits and antisocial online behaviours. Pers Individ Dif 2019 Jul;144:40-55. [doi: 10.1016/j.paid.2019.02.027]
31. Oliver E, Coates A, Bennett JM, Willis ML. Narcissism and intimate partner violence: a systematic review and meta-analysis. Trauma Violence Abuse 2024 Jul;25(3):1871-1884. [doi: 10.1177/15248380231196115] [Medline: 37702183]
32. Zhang C, Liu Y, Guo X, Liu Y, Shen Y, Ma J. Digital cognitive behavioral therapy for insomnia using a smartphone application in China: a pilot randomized clinical trial. JAMA Netw Open 2023 Mar 1;6(3):e234866. [doi: 10.1001/jamanetworkopen.2023.4866] [Medline: 36972049]
33. Zeng Y, Guo Y, Li L, et al. Relationship between patient engagement and depressive symptoms among people living with HIV in a mobile health intervention: secondary analysis of a randomized controlled trial. JMIR Mhealth Uhealth 2020 Oct 29;8(10):e20847. [doi: 10.2196/20847] [Medline: 33118956]
34. Guo Y, Li Y, Yu C, et al. Long-term effects of a social media-based intervention (Run4Love) on depressive symptoms of people living with HIV: 3-year follow-up of a randomized controlled trial. J Med Internet Res 2022 Jun 28;24(6):e36809. [doi: 10.2196/36809] [Medline: 35763324]
35. Campbell ANC, Rieckmann T, Pavlicova M, et al. Culturally tailored digital therapeutic for substance use disorders with urban Indigenous people in the United States: a randomized controlled study. J Subst Use Addict Treat 2023 Dec;155:209159. [doi: 10.1016/j.josat.2023.209159] [Medline: 37690525]
36. Bernal G, Jiménez-Chafey MI, Domenech Rodríguez MM. Cultural adaptation of treatments: a resource for considering culture in evidence-based practice. Prof Psychol Res Pr 2009;40(4):361-368. [doi: 10.1037/a0016401]
37. Barrera M, Castro FG, Strycker LA, Toobert DJ. Cultural adaptations of behavioral health interventions: a progress report. J Consult Clin Psychol 2013 Apr;81(2):196-205. [doi: 10.1037/a0027085] [Medline: 22289132]
38. Wingood GM, DiClemente RJ. The ADAPT-ITT model: a novel method of adapting evidence-based HIV Interventions. J Acquir Immune Defic Syndr 2008 Mar 1;47 Suppl 1:S40-S46. [doi: 10.1097/QAI.0b013e3181605df1] [Medline: 18301133]

39. Lindegaard T, Brohede D, Koshnaw K, Osman SS, Johansson R, Andersson G. Internet-based treatment of depressive symptoms in a Kurdish population: a randomized controlled trial. J Clin Psychol 2019 Jun;75(6):985-998. [doi: 10.1002/jclp.22753] [Medline: 30702758]
40. Silva MA, Jaramillo Y, Paris M Jr, Añez-Nava L, Frankforter TL, Kiluk BD. Changes in DSM criteria following a culturally-adapted computerized CBT for Spanish-speaking individuals with substance use disorders. J Subst Abuse Treat 2020 Mar;110:42-48. [doi: 10.1016/j.jsat.2019.12.006] [Medline: 31952627]
41. Añez LM, Paris M, Bedregal LE, Davidson L, Grilo CM. Application of cultural constructs in the care of first generation Latino clients in a community mental health setting. J Psychiatr Pract 2005 Jul;11(4):221-230. [doi: 10.1097/00131746-200507000-00002] [Medline: 16041232]
42. Añez LM, Silva MA, Paris M, Bedregal LE. Engaging Latinos through the integration of cultural values and motivational interviewing principles. Prof Psychol Res Pract 2008;39(2):153-159. [doi: 10.1037/0735-7028.39.2.153]
43. Yeung A, Martinson MA, Baer L, et al. The effectiveness of telepsychiatry-based culturally sensitive collaborative treatment for depressed Chinese American immigrants: a randomized controlled trial. J Clin Psychiatry 2016 Aug;77(8):e996-e1002. [doi: 10.4088/JCP.15m09952] [Medline: 27561153]
44. Sarfraz A, Siddiqui S, Galante J, Sikander S. Feasibility and acceptability of an online mindfulness-based intervention for stress reduction and psychological wellbeing of university students in Pakistan: a pilot randomized controlled trial. Int J Environ Res Public Health 2023 Apr 14;20(8):5512. [doi: 10.3390/ijerph20085512] [Medline: 37107796]
45. Zemestani M, Fazeli Nikoo Z. Effectiveness of mindfulness-based cognitive therapy for comorbid depression and anxiety in pregnancy: a randomized controlled trial. Arch Womens Ment Health 2020 Apr;23(2):207-214. [doi: 10.1007/s00737-019-00962-8] [Medline: 30982086]
46. Spruill TM, Friedman D, Diaz L, et al. Telephone-based depression self-management in Hispanic adults with epilepsy: a pilot randomized controlled trial. Transl Behav Med 2021 Jul 29;11(7):1451-1460. [doi: 10.1093/tbm/ibab045] [Medline: 33963873]
47. Zhang X, Li Y, Wang J, et al. Effectiveness of digital guided self-help mindfulness training during pregnancy on maternal psychological distress and infant neuropsychological development: randomized controlled trial. J Med Internet Res 2023 Feb 10;25:e41298. [doi: 10.2196/41298] [Medline: 36763452]
48. Benjet C, Zainal NH, Albor Y, et al. A precision treatment model for internet-delivered cognitive behavioral therapy for anxiety and depression among university students: a secondary analysis of a randomized clinical trial. JAMA Psychiatry 2023 Aug 1;80(8):768-777. [doi: 10.1001/jamapsychiatry.2023.1675] [Medline: 37285133]
49. Vaca FE, Dziura J, Abujarad F, et al. Use of an automated bilingual digital health tool to reduce unhealthy alcohol use among Latino emergency department patients: a randomized clinical trial. JAMA Netw Open 2023 May 1;6(5):e2314848. [doi: 10.1001/jamanetworkopen.2023.14848] [Medline: 37219901]
50. Zhou ES, Ritterband LM, Bethea TN, Robles YP, Heeren TC, Rosenberg L. Effect of culturally tailored, internet-delivered cognitive behavioral therapy for insomnia in Black women: a randomized clinical trial. JAMA Psychiatry 2022 Jun 1;79(6):538-549. [doi: 10.1001/jamapsychiatry.2022.0653] [Medline: 35442432]
51. Javier JR, Aguiling W, Cunanan P, et al. Short-term outcomes from a pilot randomized controlled trial evaluating a virtual culturally adapted parenting intervention among Filipino parents of school-age children. Cultur Divers Ethnic Minor Psychol 2025;31(1):124-137. [doi: 10.1037/cdp0000616]
52. Bernal G, Bonilla J, Bellido C. Ecological validity and cultural sensitivity for outcome research: issues for the cultural adaptation and development of psychosocial treatments with Hispanics. J Abnorm Child Psychol 1995 Feb;23(1):67-82. [doi: 10.1007/BF01447045] [Medline: 7759675]
53. Owen A, Brown B, Deragon K, et al. Developing an integrated cognitive behavioral therapy and motivational interviewing intervention for African Americans with mild cognitive impairment: a pilot study. Pract Innov (Wash D C) 2022;7(3):241-254. [doi: 10.1037/pri0000182]
54. Lindegaard T, Seaton F, Halaj A, et al. Internet-based cognitive behavioural therapy for depression and anxiety among Arabic-speaking individuals in Sweden: a pilot randomized controlled trial. Cogn Behav Ther 2021 Jan;50(1):47-66. [doi: 10.1080/16506073.2020.1771414] [Medline: 32603632]
55. Yamaguchi S, Ojio Y, Ando S, et al. Long-term effects of filmed social contact or internet-based self-study on mental health-related stigma: a 2-year follow-up of a randomised controlled trial. Soc Psychiatry Psychiatr Epidemiol 2019 Jan;54(1):33-42. [doi: 10.1007/s00127-018-1609-8] [Medline: 30315333]
56. Ellis K, Hosny N, Miller-Graff L. User experiences of a culturally adapted web-based intervention for posttraumatic stress disorder in Egypt: a qualitative study. Psychotherapy (Chic) 2022 Mar;59(1):13-25. [doi: 10.1037/pst0000429] [Medline: 35175092]
57. Sun S, Lin D, Goldberg S, et al. A mindfulness-based mobile health (mHealth) intervention among psychologically distressed university students in quarantine during the COVID-19 pandemic: a randomized controlled trial. J Couns Psychol 2022 Mar;69(2):157-171. [doi: 10.1037/cou0000568] [Medline: 34264696]
58. Jacobs P, Estrada YA, Tapia MI, et al. Familias Unidas for high risk adolescents: study design of a cultural adaptation and randomized controlled trial of a U.S. drug and sexual risk behavior intervention in Ecuador. Contemp Clin Trials 2016 Mar;47:244-253. [doi: 10.1016/j.cct.2016.01.014] [Medline: 26850901]

59. Barrera MJ, Berkel C, Castro FG. Directions for the advancement of culturally adapted preventive interventions: local adaptations, engagement, and sustainability. Prev Sci 2017 Aug;18(6):640-648. [doi: 10.1007/s11121-016-0705-9]
60. Barrera AZ, Wickham RE, Muñoz RF. Online prevention of postpartum depression for Spanish- and English-speaking pregnant women: a pilot randomized controlled trial. Internet Interv 2015 Sep 1;2(3):257-265. [doi: 10.1016/j.invent.2015.06.002] [Medline: 26273567]
61. Zemestani M, Hosseini M, Petersen JM, Twohig MP. A pilot randomized controlled trial of culturally-adapted, telehealth group acceptance and commitment therapy for Iranian adolescent females reporting symptoms of anxiety. J Contextual Behav Sci 2022 Jul;25:145-152. [doi: 10.1016/j.jcbs.2022.08.001]
62. Spanhel K, Burdach D, Pfeiffer T, et al. Effectiveness of an internet-based intervention to improve sleep difficulties in a culturally diverse sample of international students: a randomised controlled pilot study. J Sleep Res 2022 Apr;31(2):e13493. [doi: 10.1111/jsr.13493] [Medline: 34549852]

Abbreviations

DMHI: digital mental health intervention
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: Prospective Specific Evaluation of Reviews
RCT: randomized controlled trial
WEIRD: western, educated, industrialized, rich, democratic

Edited by J Torous; submitted 14.Jul.2025; peer-reviewed by RP Ghosh; revised version received 20.Nov.2025; accepted 21.Nov.2025; published 28.Jan.2026.

Please cite as:
Tandon T, Biswas R, Meteier Q, Daher K, Khaled OA, Meyer B, Berger T, Gupta R, Soelch CM
Retention and Engagement in Culturally Adapted Digital Mental Health Interventions: Systematic Review of Dropout, Attrition, and Adherence in Non-Western, Educated, Industrialized, Rich, Democratic Settings
JMIR Ment Health 2026;13:e80624
URL: https://mental.jmir.org/2026/1/e80624
doi:10.2196/80624

Remote Measurement-Based Care Interventions for Mental Health: Systematic Review and Meta-Analysis

Felix Machleid[1,2*], MPH, MD; Twyla Michnevich[1*], MD; Leu Huang[3], MD; Louisa Schröder-Frerkes[4], MD; Caspar Wiegmann[5,6], MD; Toni Muffel[6], MSc; Jakob Kaminski[1,2,6], MD

[1]Department of Psychiatry and Neurosciences, Charité - Universitätsmedizin Berlin, Charitéplatz 1, Berlin, Germany

[2]Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin, Germany

[3]Department of Infectious Diseases and Respiratory Medicine, Charité Campus Virchow-Klinikum, Charité - Universitätsmedizin Berlin, Berlin, Germany

[4]Clinics for Psychiatry and Psychotherapy, Clinics at the Vivantes Klinikum am Urban, Berlin, Germany

[5]Clinic for Psychiatry and Psychotherapy, Clinics at Theodor-Wenzel-Werk e.V., Berlin, Germany

[6]Recovery Cat GmbH, Berlin, Germany

[*]these authors contributed equally

Corresponding Author:
Felix Machleid, MPH, MD
Department of Psychiatry and Neurosciences, Charité - Universitätsmedizin Berlin, Charitéplatz 1, Berlin, Germany

Abstract

Background: Poor management of mental health conditions leads to reduced adherence to treatment, prolonged illness, unnecessary rehospitalization, and a significant financial burden to the health care system. Recognizing this, ecological momentary assessment (EMA) and remote measurement-based care (RMBC) interventions have emerged as promising strategies to address gaps in current care systems. They provide a convenient means to continuously monitor patient-reported outcomes, thereby informing clinical decision-making and potentially improving outcomes such as psychopathology, relapse, and quality of life.

Objective: This systematic review and meta-analysis aims to comprehensively appraise and analyze the existing evidence on the use of EMA and RMBC for people living with mental illness.

Methods: The study was conducted according to PRISMA-P (Preferred Reporting Items for Systematic Review and Explanation Meta-Analysis Protocols) guidelines and preregistered with the PROSPERO systematic review registry. A comprehensive search was conducted in 4 online databases using Medical Subject Headings terms related to mental disorders and digital technologies. Studies were included if they included adults with a formally diagnosed mental disorder and measured symptoms using EMA or RMBC. Studies were independently reviewed by subgroups of authors, and data were extracted focusing on symptom-focused or disease-specific outcomes, relapse, recovery-focused outcomes, global functioning, quality of life, and acceptability of the intervention. We performed a descriptive analysis of demographic variables and a meta-analysis of randomized controlled trials (RCTs). Risk of bias was assessed using the Cochrane risk-of-bias tool for randomized trials version 2 (RoB-2).

Results: The systematic review included 103 studies, of which 15 used RMBC. Of these, 9 were RCTs that were meta-analyzed. RMBC interventions varied in effectiveness, generally showing small but significant effects on symptom-specific outcomes, with notable effects on mania symptoms and empowerment. The mean adherence rate across studies to all tracking items was 74.5% (SD 13.98; n=38). More prompts per day, but not more items per prompt, were associated with lower adherence. Adverse effects were infrequently reported and included technical problems and psychological distress. Concerns about bias were raised, particularly regarding participants' awareness of the interventions and potential deviations from the intended protocols.

Conclusions: Although RMBC shows growing potential in improving and tailoring psychiatric care to individual needs, the evidence of its clinical effectiveness is still limited. However, we found potential effects on mania symptoms and empowerment. Overall, there were only a few RCTs with formal psychiatric diagnoses to be included in our analyses, and these had moderate risks of bias. Future studies assessing RMBC's effectiveness and long-term efficacy with larger populations are needed.

Trial Registration: PROSPERO CRD42022356176; https://www.crd.york.ac.uk/PROSPERO/view/CRD42022356176

International Registered Report Identifier (IRRID): RR2-10.1371/journal.pone.0297929

(JMIR Ment Health 2026;13:e63088) doi:10.2196/63088

KEYWORDS

ecological momentary assessment; RMBC; remote measurement-based care; MMH; mobile mental health; systematic review; meta-analysis; relapse; quality of life; mental disorder; psychological; psychiatric

Introduction

Mental health disorders have one of the highest global burdens of disease [1] and are difficult to manage. Hurdles include subjective symptom reporting [2,3], memory biases [4], complex treatment dynamics, and suboptimal coordination during transitions between inpatient and outpatient care settings [5]. Additionally, short and infrequent outpatient appointments contribute to the loss of essential information about symptom progression and treatment side effects [6]. This increases the risk of reduced treatment adherence, worsening conditions, preventable readmissions, and higher health care costs [7,8].

In response to these challenges, there has been a notable increase in the development of diagnostic and therapeutic mobile mental health (MMH) technologies [9-12]. One prominent application of MMH is remote measurement-based care (RMBC), which involves the asynchronous assessment of patient-reported outcomes outside of clinical visits. These assessments can then be used for clinical decision-making and triage purposes [13,14]. Apart from traditional retrospective patient-reported outcome assessment methods, such as validated self-report questionnaires, there is growing interest in ambulatory and diary approaches. These methods, collectively known as ecological momentary assessment (EMA), capture real-time, in situ data on patients' symptoms and well-being [15]. With advancements in technology, EMA has evolved to allow self-reporting of symptoms via the internet or mobile platforms, including web or online, SMS text messaging, or phone call–based systems [16]. Passive or sensor data integration further enhances the richness of this approach by capturing objective behavioral and physiological indicators in real-world settings, complementing the subjective self-reports provided by patients [13,14]. While RMBC and EMA share similarities in leveraging technology to enhance the understanding and treatment of mental health issues through continuous care, the literature does not always make a clear distinction between the two, often seeing them as part of a continuum in advancing personalized health monitoring and intervention.

Research has consistently demonstrated the benefits of RMBC, in that it may improve clinical outcomes and improve treatment adherence [17-19]. For example, a study involving 6424 participants diagnosed with various psychiatric conditions revealed that providing continuous feedback to therapists on symptom progression was associated with a 2-fold increase in therapeutic effects related to individual functioning, symptom load, interpersonal relationships, and social role performance [20]. Additionally, RMBC has been associated with faster remission rates than standard treatment approaches [21,22] and reduced missed outpatient appointments [21,22]. Moreover, RMBC enables clinicians to make timely and effective adjustments to treatment plans. Patients have reported finding RMBC valuable [23]. RMBC also showed potential to enhance doctor-patient communication and increase treatment motivation [24,25].

Despite the potential benefits, the integration of asynchronous measurement-based care (MBC) using digital solutions remains limited in clinical practice due to time constraints, workflow integration issues, and uncertainties about interpreting and using the data effectively [26]. While MMH technology companies develop extensive solutions, their scientific evaluation often lacks the depth seen in university settings, presenting a significant dissemination barrier for health care providers and insurers. Conversely, the proliferation of MMH technologies has led to numerous pilot and feasibility studies on RMBC systems by clinical research teams, which typically suffer from academic research limitations such as insufficient power and bias reduction strategies, resulting in incoherent and scattered evidence.

Often, the effectiveness of RMBC technologies is difficult to interpret, as they are often integrated into complex intervention bundles with unclear causal pathways and potential confounders. For example, a review found that while feedback from providers improved the therapeutic relationship and promoted help-seeking behavior in young people—both of which may be viewed as proxy markers for improved long-term treatment trajectories—it did not directly impact depression outcomes [27].

Furthermore, considerable variability exists in how data obtained from RMBC are used to inform treatment decisions. Unlike most somatic pharmacotherapy, where objective laboratory measurement results with defined thresholds often lead to discrete, standardized actions (such as medication adjustments), psychiatric treatment often includes a range of potential responses to measured outcomes (primarily based on the subjective reporting) [13]. This induces variance, resulting in heterogeneity in clinical response, further complicating the determination of appropriate end points for evaluating RMBC effectiveness and posing significant challenges for isolating its specific impact within complex, multicomponent interventions.

These unanswered research questions demonstrate the significant need for regular systematic evaluations to identify overarching trends and effects, thereby facilitating the broader adoption of MBC and MMH in routine care.

A 2018 systematic review by Goldberg et al [13] synthesized existing evidence on RMBC, including 36 unique samples, of which 13 were randomized controlled trials (RCTs). While generally supportive of RMBC's potential, the review highlighted considerable methodological heterogeneity, particularly due to RMBC often being embedded within broader multicomponent interventions [27]. Only 3 studies isolated the effects of RMBC experimentally, with 1 showing greater symptom improvement in the RMBC group and 2 finding no significant differences between intervention and control groups. The feasibility and acceptability of RMBC varied across studies, with promising adherence rates reported but concerns raised regarding decreased responsiveness over time. The review identified the need for more robust evaluations to better understand the isolated clinical impact of RMBC interventions, especially when implemented as part of multicomponent interventions, highlighting the need for further research to clarify its role and potential benefits.

This systematic review and meta-analysis presents a comprehensive overview of current evidence on RMBC in psychiatric care, building upon the findings of Goldberg et al [13]. In contrast to the study by Goldberg et al [13], this study focused on patients who underwent a manualized psychiatric diagnostic assessment and actively engaged with digital tools to report their individual experiences. Specifically, we concentrate on interventions targeting disorder-specific symptoms, relapse reduction, improvement in recovery-oriented outcomes, global functioning, and quality of life. Additionally, we provide a quantitative estimate of effects via a meta-analysis.

Methods

Search Strategy and Study Selection

The study adhered to PRISMA-P (Preferred Reporting Items for Systematic Review and Explanation Meta-Analysis Protocols) guidelines (Checklist 1) [28] and was preregistered with the systematic review registry PROSPERO (CRD42022356176). The detailed protocol was published elsewhere [29]. On August 24, 2022, and during the revision on December 21, 2024, we conducted a comprehensive search across 4 online databases (PubMed, Medline, Embase, and PsychINFO) and gray literature using terms related to mental disorders, psychological distress, MBC, and digital technologies (Table S1 in Multimedia Appendix 1).

Inclusion and exclusion criteria were defined by using the PICOS (population, intervention, comparison, outcome, and study) framework (Table S2 in Multimedia Appendix 1) [30]. Studies were included if they (1) targeted adults (≥18 y) diagnosed with a mental health disorder according to the *International Classification of Diseases* (*ICD*) or *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*) [31,32]; (2) implemented interventions centered on the digital assessment of self-reported symptoms or well-being factors to guide clinical decision-making or treatment planning; and (3) reported quantitative outcomes related to symptoms, recovery, functioning, or quality of life. Eligible studies had to be published in English or German. No restrictions were placed on comparator conditions to broaden the evidence base. While the systematic review included randomized and nonrandomized studies, the meta-analysis was restricted to RCTs only.

Data Extraction

The systematic extraction process was described in the study protocol [29]. Subgroups of authors (T Michnevich and LPSF; FM and LH; JK, CW, and T Muffel) independently reviewed the abstracts and full texts, resolving any discrepancies through group consensus. A comprehensive dataset, encompassing study identification (author, year of publication, DOI, and URL), population (eg, the number of cases and controls, diagnosis, age, gender, and years of preuniversity education), tracking (eg, mode, number and content of items, and frequency), and study characteristics (eg, design, hypotheses, study site, duration, randomization, postassessment period, follow-up, outcomes, and response rate), was extracted. Outcomes were systematically grouped into 6 predefined categories: symptom-focused or disease-specific outcomes, relapse, recovery-focused outcomes (in particular, empowerment), functioning or global functioning, quality of life, and acceptability.

Data Synthesis and Statistical Analysis

RStudio statistical software (version 2023.09.1+494; Posit PBC) [33] was used for statistical analysis. Demographic variables were descriptively analyzed by calculating means and SDs. A linear regression model was used to explore the impact of daily prompt frequency and the number of tracking items on participant response rate.

Frequentist Meta-Analysis

Data Synthesis

Random-effects meta-analyses were performed using the *metafor* package (version 4.6‐0) [34]. Outcomes were meta-analyzed when at least 3 studies (n>2) reported comparable results. Only instruments with evidence of construct validity or sufficient correlation with other instruments were included. When multiple instruments within a study measured the same construct, the outcome most commonly reported across studies was included to ensure comparability.

We included all measures of psychopathology, even if they were not disease-specific, for example, measures of depression in a sample of patients with psychosis. This approach recognizes the transdiagnostic nature of symptoms and prioritizes symptoms over diagnoses. The full list of constructs and outcomes can be found in Table S3 in Multimedia Appendix 1.

Intention-to-treat data were used for analyses where available. Where outcomes were reported as medians and IQRs, means and SDs were estimated using median-based imputation [35]. If only SEs were reported, SDs were calculated [36]. For trials that reported outcome data at multiple follow-up points, data from the time point immediately after the end of the intervention were used.

Effect sizes for continuous measures were expressed as standardized mean differences (SMDs), calculated by using the pooled SD of the interventions. SMDs are presented as values of Hedges *g*, along with their 95% CI.

Assessment of Heterogeneity

Heterogeneity between studies was evaluated using the I^2 statistic and by visual inspection of the forest plots. Heterogeneity was defined as very low, low, medium, and high heterogeneity when I^2 values were <25%, 25% to <50%, 50% to <75%, and ≥75%, respectively [37].

Assessment of Publication Bias

Publication bias was evaluated by visual inspection of funnel plots assessing the symmetry of effect size distributions relative to SEs.

Bayesian Meta-Analysis

To complement the frequentist approach and to better account for uncertainty due to the small number of studies, a random-effects Bayesian meta-analysis was performed using the *bayesmeta* package (version 2.21) [38-41]. Posterior distributions for the overall effect and heterogeneity parameters were estimated via Markov Chain Monte Carlo simulations [42]. Given the paucity of literature on RMBC interventions and the lack of prior knowledge, we used weakly informative priors μ=0 and σ=4 [43,44]. The prior for between-study heterogeneity τ=0.5 was set using a half-normal distribution [45]. Results are presented through marginal posterior density plots, illustrating uncertainty around overall effects and heterogeneity.

Risk of Bias

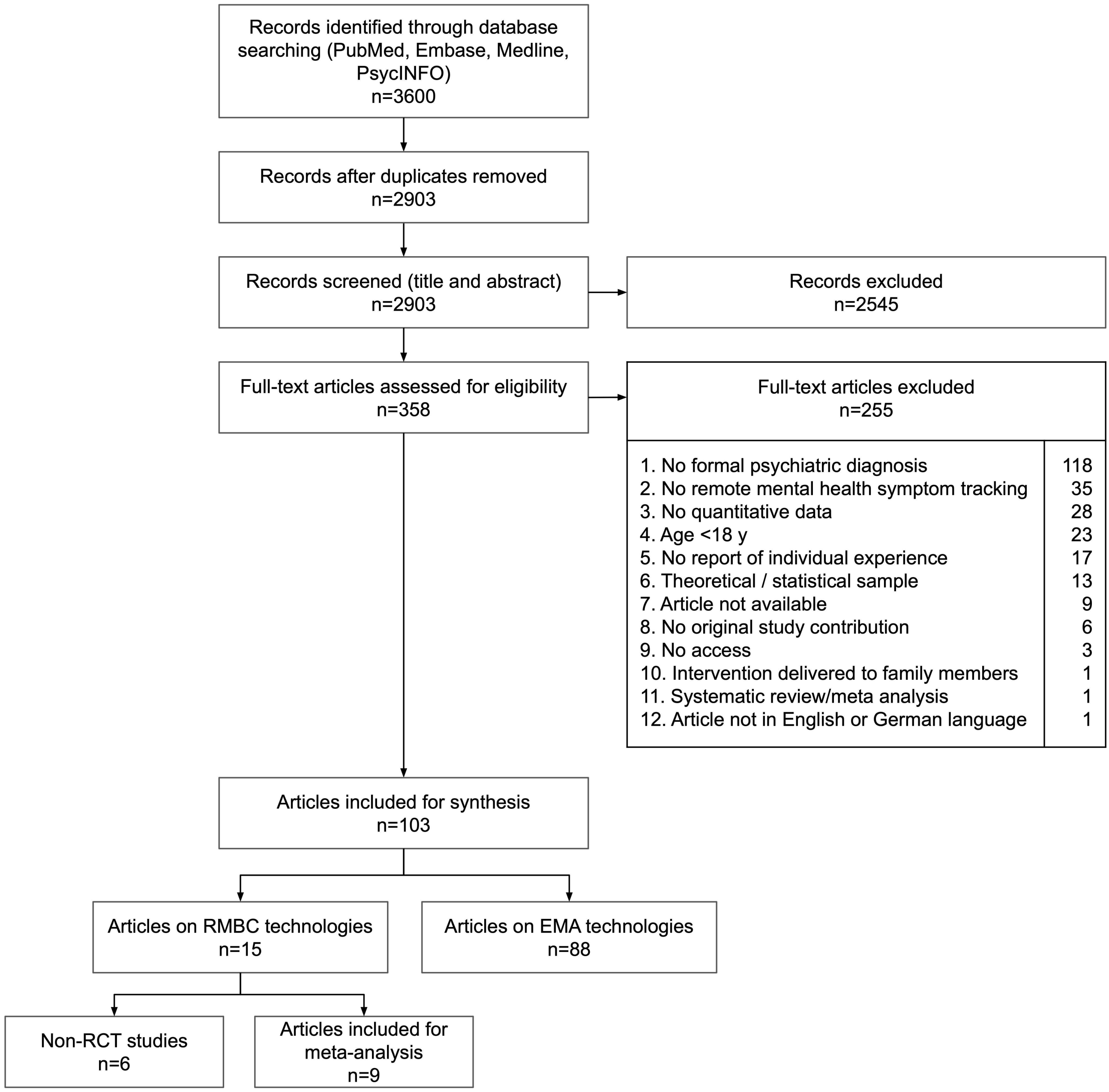
The risk of bias was evaluated independently by 2 researchers (FM and T Michnevich), using the Cochrane risk-of-bias tool for randomized trials version 2 (RoB 2) [37]. The researchers assessed potential biases across the 5 domains of the ROB 2 tool: randomization process, effect of assignment to intervention, missing outcome data, measurement of outcome, and selection of reported results [46]. Studies were classified as low risk when all domains were deemed low risk. A study was considered to have some concern if any of the domains raised concerns. The overall risk was designated as high if at least one domain was rated as high risk. Disagreements were resolved through discussion to reach consensus.

Results

Selection of Studies

The database search (N=3599) yielded 2902 records after deduplication (Table S1 in Multimedia Appendix 1), which were screened by title and abstract. Of the 357 records that qualified for full-text analysis, 254 records were excluded for not meeting the inclusion criteria (Figure 1). The most common reason for exclusion (n=118) was the lack of a formal psychiatric diagnosis, using either the *ICD* or *DSM*. The systematic review includes a final sample of 103 studies representing 109 unique samples.

Figure 1. Flowchart of the search and selection process. EMA: ecological momentary assessment; RCT: randomized controlled trial; RMBC: remote measurement-based care.



The systematic review revealed that 15 studies examined RMBC sensu stricto (using data to support clinical decision-making or treatment planning), of which the 9 RCT studies were used for meta-analysis. The other studies (n=88) matched the definition of EMA, whereby technologies were also used to collect mental health data remotely in real time, but the data did not have a significant impact on treatment.

Study Characteristics—Overall Sample

Of the 103 studies that were systematically analyzed, 41 contained healthy or diagnosis-matched control groups. Individual samples (n=109) varied due to overlapping or related datasets (Table S4 in Multimedia Appendix 1). Across the studies, the mean sample size of cases was 80.33 (SD 105.17), with an average participant age of 40.38 (SD 7.70) years and a mean study-level proportion of 56.79% (SD 22.43) female participants. For the studies including control groups, the average sample size was 62.49 (SD 71.70) controls with an average age of 41.56 (SD 6.92) years and a mean proportion of 55.04% (SD 20.42) female participants. Educational attainment was reported through various metrics, the most frequent being total years of education, which averaged at 13.62 (SD 1.22) for cases and 13.52 (SD 1.25) for control samples. The most common population was participants with schizophreniform disorders (n=21, 20.4%; ie, F2x diagnoses), followed by bipolar disorder (n=19, 18.4%). Most studies (n=70) used digital prompts formulated by the study team, while 15 studies used validated questionnaires. Six studies used a combination of both methods, and 7 studies incorporated prompts that were individually generated by the participants themselves. The predominant mode of remote data collection involved smartphones or mobile phones owned by the participants themselves. Adherence to data entries was mainly measured by the percentage of total measurements entered by the participants.

Study Characteristics—RMBC Studies

Table 1 and Tables S5 and S6 in Multimedia Appendix 1 report information extracted from the RMBC studies. Chermahini et al [47] reported demographic information jointly for the case and control groups, which is why there are only 8 unique samples for the demographic data, not 9. Across the 8 samples, the intervention groups had a mean of 66 (SD 58.98) cases, with a mean age of 38.53 (SD 7.14) years and a mean proportion of 54.75% (SD 17.93) female participants across studies. Comparably, the control groups (n=8) had a mean of 58.63 (SD 60.38) participants, with a mean age of 39.93 (SD 7.08) years and a mean proportion of 52.89% (SD 15.25) female participants across studies. Most of the studies (n=6) did not report on education, and the remaining studies used varying measures. Three studies included patients with schizophreniform disorders [48-50]; others included patients with bipolar disorder (n=2), borderline personality disorder (n=1), generalized anxiety disorder (n=1), and a range of different diagnoses or transdiagnostic symptoms (n=2). On the patient side, the majority of RMBC systems (n=6) were mobile phone- or smartphone-based. All interventions consisted of self-administered symptom tracking along with additional formalized (eg, psychotherapy) or informal psychiatric or psychotherapeutic support.

Table . Characteristics of remote measurement-based care (RMBC) studies included in the meta-analysis.

Study	Participants, n (%)		Psychiatric disorder	Setting	Treatment	Control	RMBC device	Isolated RMBC	Response rate (%)	ITT[a]	Clinical effectiveness or efficacy
	Treatment group	Control group									
Cullen et al (2020) [48]	• 28 (68.3)	13 (31.7)	• Schizophrenia and schizoaffective disorder	Hospital-based community psychiatry program	• Self-administered assessment and automated intervention with additional support by health care providers	• TAU[b]	Mobile phone	Yes	N/A[c]	No	Positive
Chermahini et al (2024) [47]	• Group 1: 45 (46.9) • Group 2: 51 (53.1)	N/A	• Generalized anxiety disorder	Psychiatric outpatient clinics	• Treatment 1: electronic cognitive behavioral therapy including homework with personalized feedback from health care providers	• Treatment 2: weekly asynchronous online mental health check-in questions with personalized clinical feedback	Online platform (treatment 1), asynchronous messaging system (online psychotherapy tool; treatment 2)	Yes	N/A	Yes	Positive

Study	Participants, n (%)		Psychiatric disorder	Setting	Treatment	Control	RMBC device	Isolated RMBC	Response rate (%)	ITT[a]	Clinical effectiveness or efficacy
	Treatment group	Control group									
Ebert et al (2013) [51]	• 21 (100)	N/A	• Affective disorders • Neurotic, stress-related, and somatoform disorders • Behavioral syndromes associated with physiological disturbances and physical factors • Disorders of adult personality and behavior	Psychiatric inpatient and outpatient treatment	• In-person psychotherapy • Internet-based maintenance therapy with additional coaching support	• In-person psychotherapy	Not specified (web-based)	No	N/A	Yes	Positive
Faurholt-Jepsen et al (2020) [49]	• 85 (65.9)	44 (34.1)	• Bipolar disorder	Psychiatric outpatient clinics		• TAU	Smartphone	No	72.6	Yes	No significant difference

Study	Participants, n (%)		Psychiatric disorder	Setting	Treatment	Control	RMBC device	Isolated RMBC	Response rate (%)	ITT[a]	Clinical effectiveness or efficacy
	Treatment group	Control group									
					• Self-administered symptom assessment with additional support from health care professionals						
Faurholt-Jepsen et al (2021) [50]	• 47 (48.0)	51 (52.0)	• Bipolar disorder	Psychiatric outpatient clinics	• Self-administered symptom assessment with additional support from health care professionals	• TAU	Smartphone	No	80.6	Yes	No significant difference
Gallinat et al (2021) [52]	• 12 (48.0)	13 (52.0)	• Schizophrenia spectrum disorder	Research clinic		• TAU	Smartphone or mobile phone	No	70.7	N/A	Positive

Study	Participants, n (%)		Psychiatric disorder	Setting	Treatment	Control	RMBC device	Isolated RMBC	Response rate (%)	ITT[a]	Clinical effectiveness or efficacy
	Treatment group	Control group									
					• Collaborative care • Self-administered symptom assessment with additional support from health care professionals						
Laursen et al (2021) [36]	• 42 (53.8)	36 (46.2)	• Borderline personality disorder	Psychiatric outpatient clinics	• In-person psychotherapy (DBT[d]) collaborative care • Self-administered symptom assessment and psychoeducation	• In-person psychotherapy (DBT) • Paper-based diary cards	Smartphone	No	N/A	Yes	Positive and negative
Lewis et al (2020) [53]	• 40 (49.4)	41 (50.6)		Outpatient mental health facility		• TAU	Smartphone	No	N/A	Yes	Positive

Study	Participants, n (%)		Psychiatric disorder	Setting	Treatment	Control	RMBC device	Isolated RMBC	Response rate (%)	ITT[a]	Clinical effectiveness or efficacy
	Treatment group	Control group									
			• Schizophrenia spectrum disorder		• Self-administered symptom assessment with additional support from health care professionals						
Spaniel et al (2015) [54]	• 74 (50.7)	72 (49.3)	• Schizophrenia and schizoaffective disorder	Psychiatric outpatient clinic	• Self-administered symptom assessment with additional support from psychiatrists	• TAU	Mobile phone to computer SMS interface	No	N/A	Yes	No significant difference

[a]ITT: intention to treat.

[b]TAU: treatment as usual.

[c]N/A: not applicable.

[d]DBT: dialectical behavior therapy.

Effectiveness of RMBC Interventions

While most studies found RMBC interventions to be effective, the others found no effects (n=3) or mixed results. Faurholt-Jepsen et al [49] found no benefit of a 9-month self-administered symptom assessment that provided patients with automated predictions of future mood states. In an exploratory subgroup analysis, patients in the intervention group were more likely to experience a relapse of depressive symptoms than patients receiving usual outpatient care.

Adverse Effects of RMBC

Three RMBC studies reported adverse or potential negative effects of the interventions. These included technical malfunctioning, psychological distress attributed to prompts [53], hospitalization within the trial period (notably considered an outcome parameter, not an adverse effect, by several other studies) [54]; and changes to patient-therapist interactions due to the new technology [36].

Frequentist Random Effects Meta-Analysis

Data related to relapse and readmission rates were inconsistent between the studies with differing periods of observation. Thus, no meta-analysis of the data was possible.

Meta-Analysis of Symptom-Focused Outcomes

Regarding psychotic symptoms (Figure 2), data from 3 studies (n=143) showed a small nonsignificant effect (SMD −0.20, 95% CI −0.53 to 0.14; *P*=.20). For depressive symptoms (Figure 3), a larger sample of 5 trials (n=423) showed a nonsignificant overall effect (SMD −0.00, 95% CI −0.37 to 0.36; *P*>.99). For manic symptoms (Figure 4), data from 4 studies (n=264) revealed a moderate to large significant effect of RMBC interventions (SMD −0.80, 95% CI −1.28 to −0.32; *P*<.001). Data from one large transdiagnostic study (Figure 5; n=400) suggested a moderate and significant effect size with an SMD of −0.29 (95% CI −0.40 to −0.17; *P*<.001). Between-study heterogeneity was low for psychotic symptoms (I^2=0%), moderate for depressive symptoms (I^2=72%), and moderate for manic symptoms (I^2=68%), suggesting varying degrees of similarity between the studies within each construct.

Figure 2. Forest plot of pooled effect on psychotic symptoms [48,52,53].

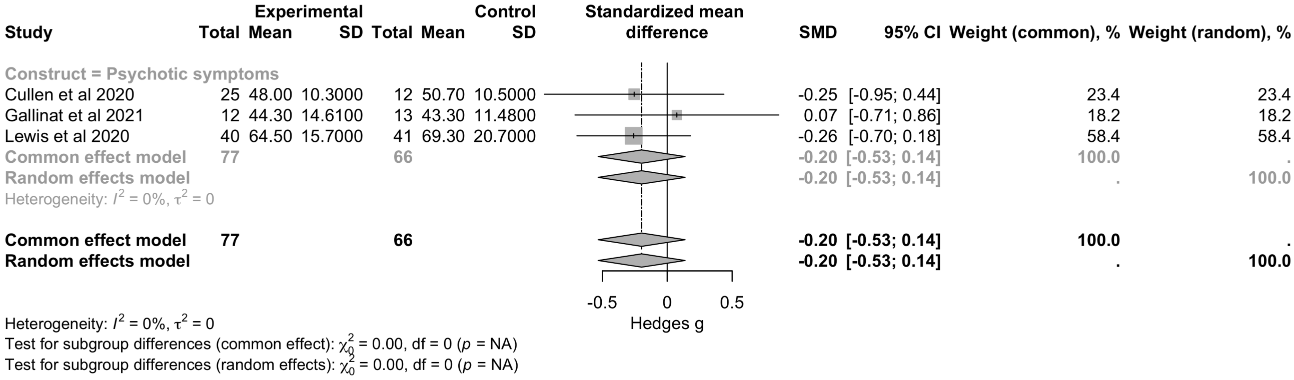


Figure 3. Forest plot of pooled effect on depressive symptoms [36,48-50,53].

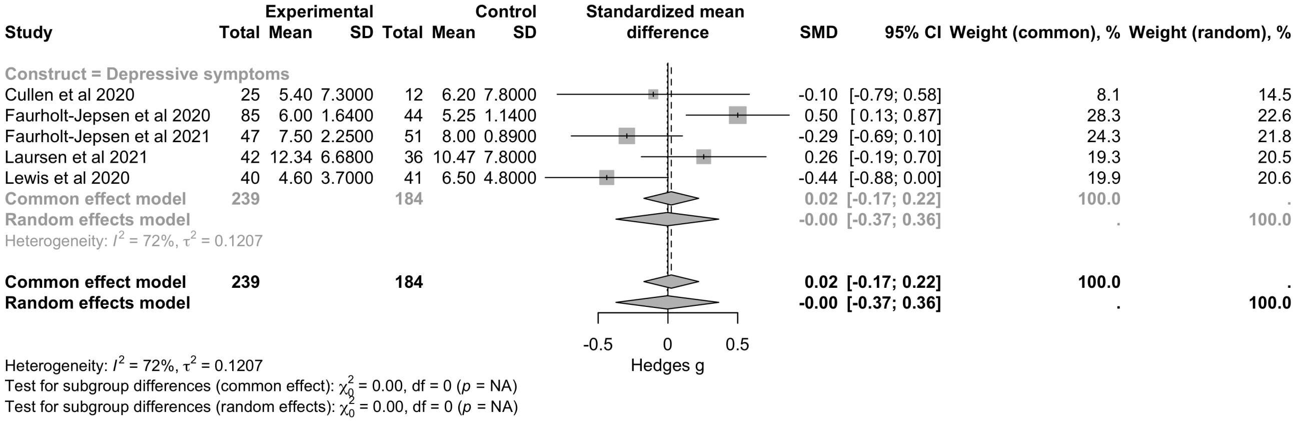


Figure 4. Forest plot of pooled effect on manic symptoms [48-50].

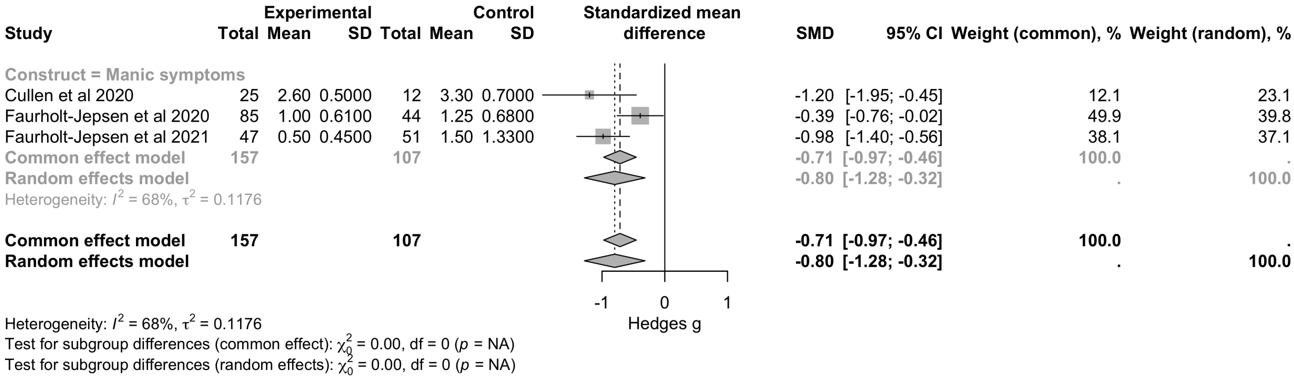
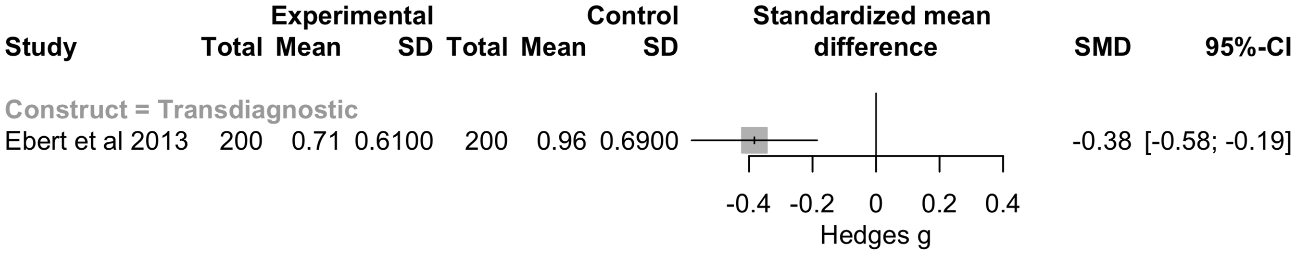


Figure 5. Forest plot of effect on transdiagnostic symptoms [51].



Meta-Analysis of Empowerment, Quality of Life, and Functioning

For the construct of empowerment and self-efficacy (Figure 6), pooled data from 3 studies (n=518) demonstrated a small-to-moderate positive effect (SMD 0.39, 95% CI 0.21 to 0.56; *P*<.001). Regarding quality of life (Figure 7), combined results from 4 studies (n=601) showed a nonsignificant effect (SMD −0.01, 95% CI −0.40 to 0.38; *P*>.99). For functioning (Figure 8), the analysis included 2 studies (n=179) and reported a nonsignificant effect (SMD −0.16, 95% CI −0.45 to 0.14; *P*>.99). Heterogeneity between the studies was zero for empowerment and functioning (I^2=0%) and high for quality of life (I^2=85%).

Figure 6. Forest plot of pooled effect on empowerment and self-efficacy [48,51,53].

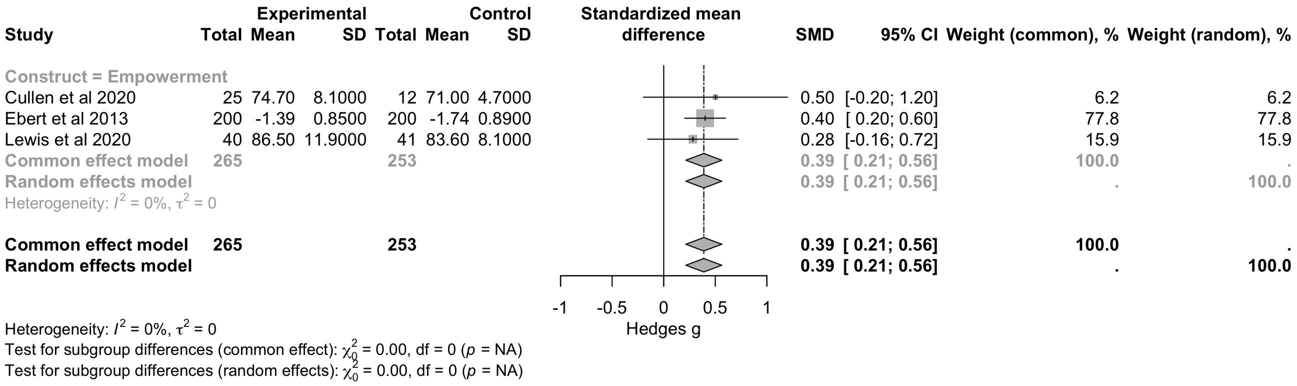


Figure 7. Forest plot of pooled effect on quality of life [36,51,53].

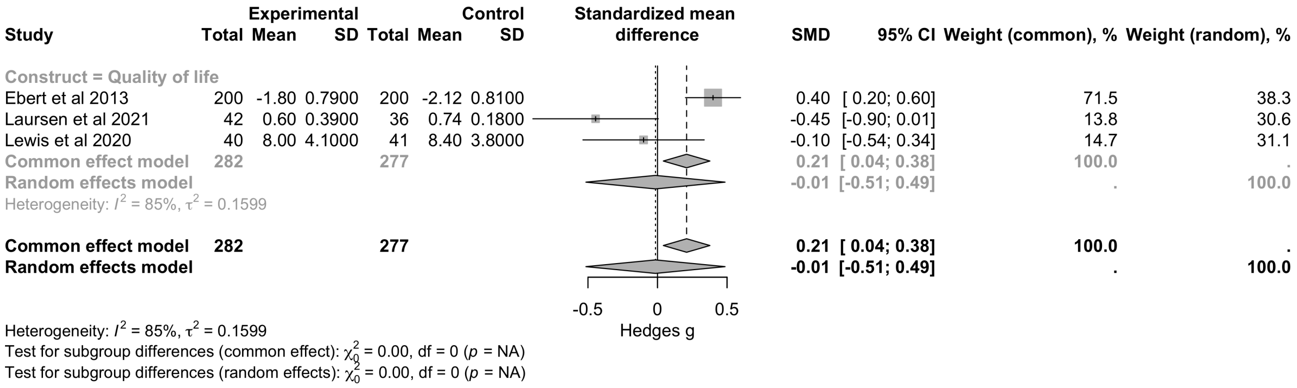
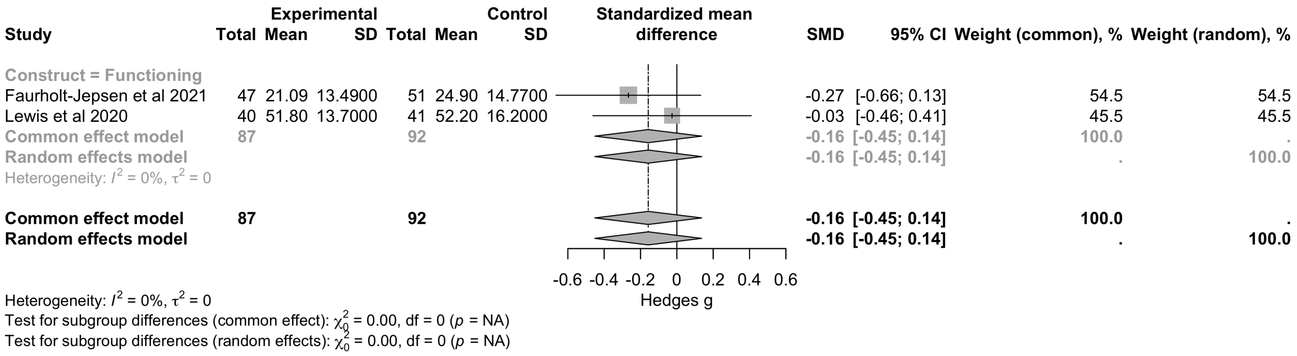


Figure 8. Forest plot of pooled effect on functioning [50,53].



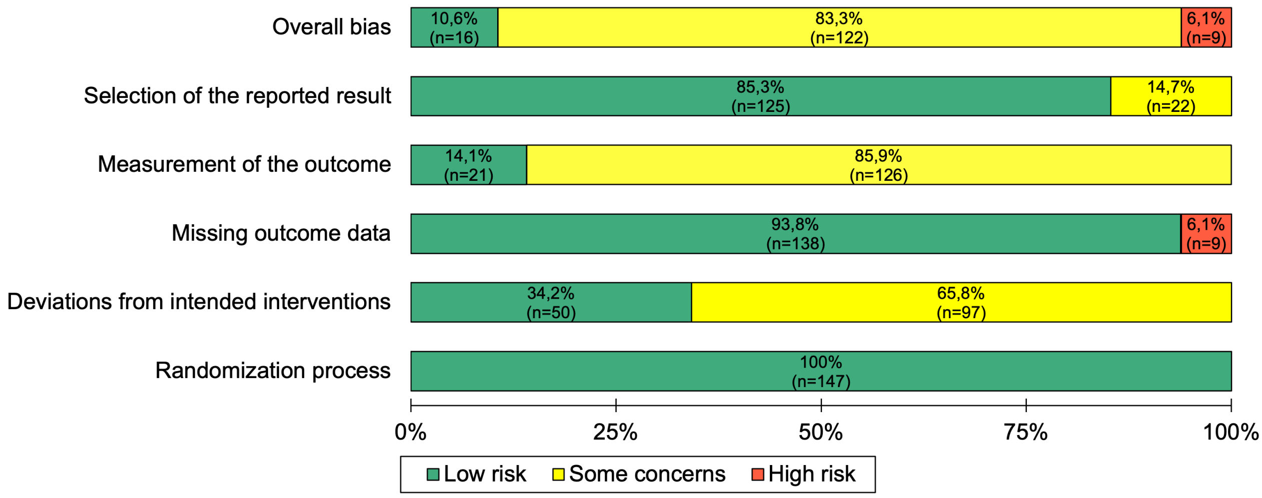
Bayesian Meta-Analysis

In general, the Bayesian meta-analysis yielded similar results to the frequentist meta-analysis, but there was only one significant result: the weighted pooled effect size with a mean estimate of −0.79 (95% CI −1.44 to −0.20), for the reduction in manic symptoms associated with RMBC interventions. There was moderate heterogeneity τ=0.36 (0.00-0.84). The prediction interval of −1.99 to 0.34 reflected moderate uncertainty in predicting new effects based on current data (Figure S7C in Multimedia Appendix 1). The effect on empowerment that was significant in the frequentist analysis showed an effect size of 0.39 and the CI crossed zero (95% CI −0.02 to 0.79). Other analyses showed nonsignificant effects on the outcomes assessed (Figures S7A,B,D-F and S8 in Multimedia Appendix 1).

Risk of Bias

The overall risk of bias indicates that the majority of outcomes were of concern to the reviewers (Figure 9). This was largely due to participants, caregivers, or assessors having been aware of the assigned digital interventions, which made it difficult to assess the outcomes, particularly since many relied on participant-reported data. Further, the effect of assignment to the intervention raised concerns about deviations from the intended interventions. Specifically, the outcomes of the Boston University Empowerment Scale in Cullen et al [48] were rated as having a high risk of bias because missing data were replaced by scale means from follow-up data. In addition, the reviewers identified a high risk of bias in the outcome data from Chermahini et al [47] due to high dropout rates and missing data. A full assessment of each outcome is provided in Figure S10 in Multimedia Appendix 1.

Figure 9. Cochrane risk of bias summary. Authors' judgments about each risk of bias item across all assessment time points.



Tracking and Adherence—Overall Sample

Typically, participants were prompted to complete questionnaires >1× per day, followed by daily EMA, with the number of items ranging between 1 and 43 (mean 13.87, SD 10.55) per session (in 16 studies, the number of items was unclear, and in 5 studies, the number of items varied). The most granular tracking data were collected by Freedman et al [55] with 128 to 136 tracking items per day, amounting to a minimum of 896 individual data points per week. Thirteen studies included additional passive data sensing such as GPS, phone usage, speech activity, ambient noise and light, and sleep activity. Forty-seven studies provided a metric of EMA or RMBC adherence with an overall mean response rate of 74.64% (SD 13.04%) to the prompts.

Association Between Tracking Frequency and Adherence

A linear regression model investigated the effect of the number of prompts per day and the number of tracking items per day on response rate (Figure 10). The model results (Table 2) showed a significant negative effect of the logarithm of the number of prompts per day (P=.02) and no effect of the number of tracking items on the response rate. Diagnostic plots showed no obvious violations of the key assumptions (Table S9 in Multimedia Appendix 1). There was no multicollinearity between the independent variables as all variance inflation factor values were less than 5. The model's overall fit was sufficiently good, given the residual plots (adjusted R^2=0.0173), but other variables may have affected the response rate. Due to omitted variable bias, the model may overestimate the effect of predictors.

Figure 10. Bubble plot visualization with the predicted probability line with 95% CI (gray area) for the response rate (%) as a function of the log-transformed number of daily prompts on the x-axis (est=−7.121, t_{26}=−2.540; *P*=.02) and the number of tracking items per prompt, represented as the bubble size (est=3.379, t_{26}=1.306; *P*=.20).

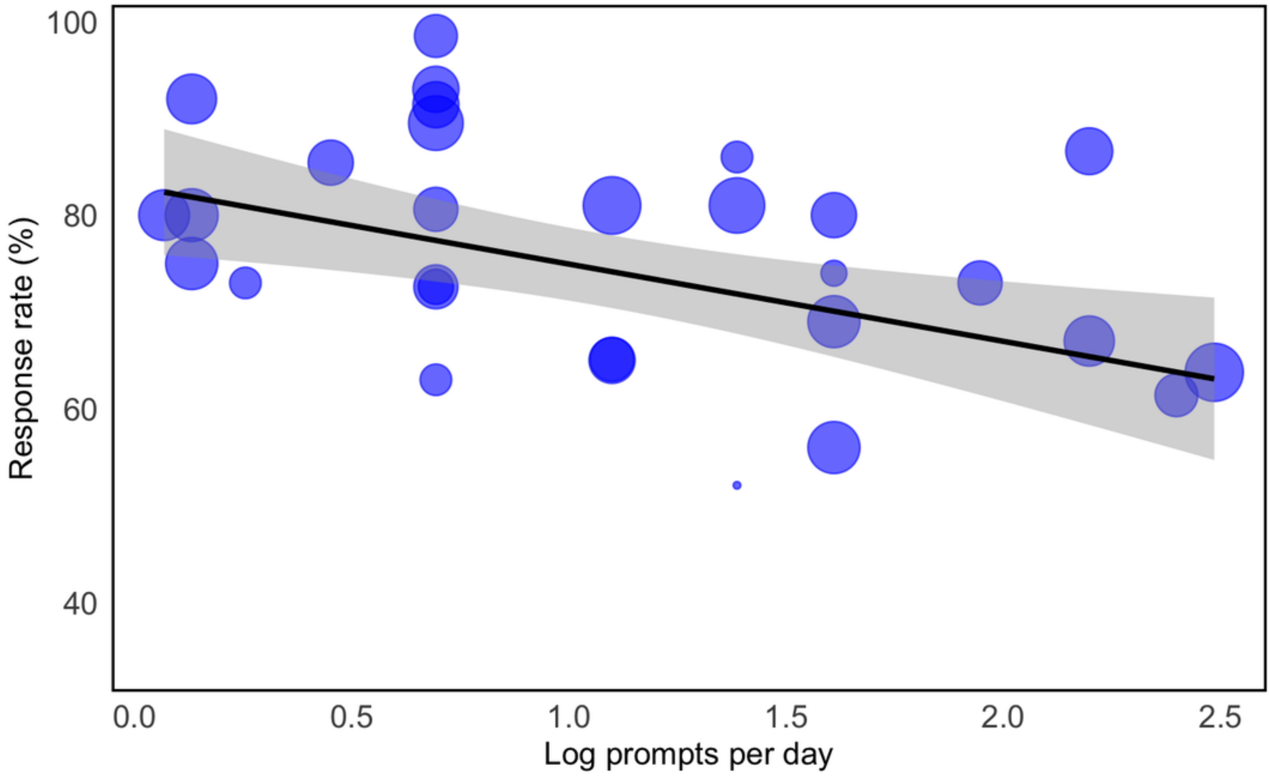


Table . Summary table of the linear regression model results[a].

Coefficients	Estimate	SE	*t* test (*df*)	*P* value	VIF[b]
(Intercept)	75.617	7.217	1,047,810.072 (26)	<.001	—[c]
Log prompts per day	−7.121	—	−2.540 (26)	.02[d]	1.04
Log tracking items	3.379	2.587	1.306 (26)	.20	1.04

[a]Linear model: response rate approximately log prompts per day + log number of tracking items per prompt.

[b]VIF: variance inflation factor.

[c]not applicable.

[d]*P*<.05.

Discussion

Principal Results

Given the widespread access to smartphone technology, the steadily advancing EMA and RMBC research, and the limited evidence through RCTs and systematic studies on interventions in mental health care, we aimed to review and evaluate the diverse literature in the field. This systematic review targeted the study design features and procedures of EMA and RMBC across psychiatric disorders. Concurrently, the meta-analysis aggregated and examined the effects of RCTs implementing RMBC interventions, focusing on outcomes pertinent to clinical efficacy and recovery-oriented outcomes.

Overall, we found compliance and retention rates for RMBC and EMA technologies to be encouraging, aligning with previous findings in broad EMA research [56,57]. We found that more prompts, but not tracking items, negatively affected the response rate. This observation corroborates the meta-analytic evidence by Vachon et al [57], who noted a positive correlation between compliance and fewer daily prompts as well as longer intervals between prompts in severe mental illnesses [50]. It also confirms findings from a systematic review by Williams et al [58], showing that higher numbers of tracking items per prompt were not associated with reduced compliance in clinical samples; in healthy individuals, however, more items were indeed associated with lower compliance. Up to 5 random EMA prompts per day have been deemed optimal for longitudinal studies [59]. However, in the context of substance use disorders, Jones et al [60] reported that compliance was not significantly impacted by the number of prompts per day or the duration of the assessment period. Our results support the evidence for severe mental illness and are in favor of longer intervals between successive evaluations to maximize, potentially influenced by the low representation of substance use disorders within our sample due to a lack of formal diagnosis.

Methodology

During the systematic review of the literature, a prominent distinction was identified between RMBC and EMA. Despite their shared aim of collecting subjective, real-time data from remote settings, they serve distinct purposes from a clinical perspective. While RMBC encompasses elements of EMA, it is directed toward informing health care decisions and interventions, for example, supporting real-time and asynchronous treatment adjustments or scheduling of visits [13,14]. For patients, both EMA and RMBC interfaces facilitate reflection on symptoms, with many offering data summaries on symptom trajectories. As the primary difference between RMBC and EMA, RMBC focuses on enabling clinicians to formulate recommendations and implement treatment adjustments based on real-time data. As a result, patients may perceive RMBC as involving closer monitoring, which in turn is subject to individual preference. Some patients may interpret RMBC as an invasion of their autonomy and privacy, while others may find comfort in the increased level of monitoring, viewing it as an additional safety measure.

In our meta-analysis of RMBC interventions, we investigated the transdiagnostic benefits of the technologies often emphasized in the literature. Therefore, the analysis considered psychopathological, cross-diagnostic constructs rather than individual diagnostic groups of participants. In addition to the well-known challenges of different design features and procedures when integrating and aggregating data from EMA and RMBC studies [13,56], this aspect may have increased heterogeneity. Overall, we did not observe clear effects of RMBC interventions for most of the constructs we analyzed, that is, depressive, psychotic symptoms, quality of life, and daily functioning. This is consistent with the results of Goldberg et al [13], who assumed a general effect of RMBC interventions but did not draw any conclusions about specific effects due to the small number of RCTs.

There was no overlap between the studies we included in our review and those by Goldberg et al [13]. This is because we only included studies with manualized psychiatric diagnostic procedures and applied a narrower definition of RMBC, emphasizing a possible clinical need including self-monitoring for decision-making and therapy planning. Goldberg et al [13] did not set a formal psychiatric diagnosis as an inclusion criterion and used a wider definition of RMBC, in particular regarding its direct effects on treatment trajectories.

Both frequentist and Bayesian meta-analyses demonstrated a significant effect on the reduction of manic symptoms when pooling data from 3 studies [48-50]. From both RCTs by Faurholt-Jepsen et al [49,50], which found no significant effect on emotional (depressive and manic) symptoms, medians and SEs were converted to means and SDs. It is generally known that the standardization of results can introduce flaws in meta-analysis [61]. Therefore, this result has to be considered with caution. Although there is scarce systematic evidence on the effects of RMBC on manic or hypomanic symptoms, there appears to be a benefit from clinical practice due to the dynamic and fluctuating nature of the symptoms and also clear recommendation on symptom tracking in guidelines [62,63]. The exploratory analysis by Faurholt-Jepsen et al [49] underlined that smartphone-based monitoring may reduce the risk of relapse of manic episodes but increase the risk of relapse of depressive episodes. This finding is underscored by a systematic review by Hennemann et al [64], who examined internet- and mobile-based tools for psychiatric aftercare and relapse prevention. They found small to moderate symptom reduction, with the best evidence for depression and anxiety [64].

In the frequentist meta-analysis, we found an effect of RMBC interventions on empowerment and self-efficacy. Of note, the largest study (n=200) by Ebert et al [51] contributed the largest weight to this result. Although validated, the instrument used includes a limited set of 5 items as a subscale of the HEALTH-49 questionnaire and has no proven correlation with the BUES [65]. These results should also be evaluated with caution in light of the results of the Bayesian meta-analysis where the effect was also detectable but the CI includes zero. Overall, self-efficacy seems a promising target for RMBC tools.

Quality of Evidence

A keyword search of the manuscripts in this systematic review and meta-analysis showed that unfortunately, none of the publications mentions the CONSORT (Consolidated Standards of Reporting Trials) checklist or its EHEALTH (Electronic Health) extension. Additionally, some essential outcome measures were missing from the vast majority of studies.

For one, adverse events were reported by only a fraction of studies. This is particularly surprising as adverse events pertaining to the technologies and treatment modalities may be easily transferable, constituting an efficient knowledge transfer and reducing potential harm to study and clinical populations. Direct adverse psychological effects of symptom tracking include anxiety or obsessiveness about choosing the “wrong” answer and an increased awareness of symptoms mentioned in questions or prompts [60], which may increase disease burden. Symptom tracking has also been found to potentially amplify symptoms or create the illusion of symptom amplification for patients and clinicians through over-reporting [66].

Indirect negative effects include feelings of guilt when tracking is missed, cognitive dissonance due to continuous confrontation with mental illness, and boredom or fatigue [67]. Symptom-tracking apps may also promote individualist models of illness that negate social determinants of health and make patients indirectly responsible for their illness if they refuse or fail to track symptoms [68]. Shared decision-making and using routine outcome monitoring collaboratively could address this concern and has been shown to increase the working alliance in mental health care [66].

Incomplete or absent adverse event reporting may be linked to the circumstance that many EMA or RMBC studies are financed by industry, and funding for further product development may be dependent upon its evaluation, constituting a potential conflict of interest. This conflict of interest has to be taken into account when evaluating the effectiveness of MMH apps.

A further concern identified within the analyzed studies is the limited reporting of adherence and response rates. This underreporting poses significant methodological and interpretive challenges. Adherence and response rates are critical indicators of the feasibility and acceptability of interventions among participants. Limited or absent reporting of these metrics hinders a full understanding of intervention effectiveness and the factors influencing user engagement. Without clear insights into adherence and response rates, it becomes difficult to determine the reliability and generalizability of study findings. A further obstacle here is the multitude of metrics used, for example, the percentage of total assessments completed [69], the percentage of days within the observation period on which assessments were completed [70], or a binary definition of compliance or noncompliance based on a cutoff of completed assessments [71]. We therefore strongly emphasize the importance of standardized adherence reporting. Minimally, authors should report the total share of assessments completed within the study population, the average percent of assessments completed per person, and factors associated with nonadherence (ie, demographics or time-varying factors [72,73]).

A common barrier to evidence synthesis that affected this research is the heterogeneity of study populations, specifically the lack of a formal psychiatric diagnosis in many study samples. About a third of the full-text articles were primarily excluded for this reason. As many of the excluded study populations most likely fulfilled *DSM-5* or *ICD-10* diagnostic criteria, this issue underscores the importance of standardized diagnostic criteria and rigorous documentation of participant characteristics in clinical research.

Strengths and Limitations

On the one hand, our inclusion criteria required a formal psychiatric diagnosis, which led to the exclusion of many studies that used EMA and RMBC technology. On the other hand, this criterion also strengthens the methodology by ensuring a higher standard of diagnostic rigor within the included studies. In addition, our focus on RCTs increases the reliability of our findings by selecting evidence from studies with robust experimental designs. The heterogeneity of terminology in studies exploring similar concepts, such as EMA or RMBC, may have led to the inadvertent omission of relevant research. To meta-analyze constructs, outcome constructs were pooled, reducing their discriminatory power, possibly leading to an underestimation or nondetection of effects [74]. Even after an extended screening period during the revision phase, only 16 studies met our inclusion criteria and were included in the analysis, highlighting a gap between progressive technological innovation and rigorous clinical validation.

Recommendations

From studying the existing evidence on EMA and RMBC for mental health care, we recommend adherence to standardized reporting guidelines such as CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth). To effectively analyze acceptability and adherence, we suggest the establishment of standards for response-rate measurement. To gain feedback on the user experience as well as the perspective of health care providers with RMBC products, mixed methods designs can provide valuable insights for challenges in implementations of such measures.

Conclusions

In conclusion, our systematic review and meta-analysis underscore the potential of RMBC interventions in enhancing the management of mental health conditions, particularly in reducing symptom severity in mania and increasing empowerment. While demonstrating promising effects on adherence and symptom-specific outcomes, the variability in intervention effectiveness and concerns about bias highlight the need for further research and refinement to optimize the implementation of RMBC within mental health care systems.

Acknowledgments

The authors want to thank their colleagues at Recovery Cat for their support and thoughtful input.

Conflicts of Interest

JK is a shareholder and managing director of Recovery Cat GmbH. TM is an employee at Recovery Cat GmbH. CW received remuneration from Recovery Cat for consulting.

Multimedia Appendix 1

Supplementary materials for the systematic review and meta-analysis, including search strategies, eligibility criteria, outcome mappings, study and intervention characteristics, forest plots, risk of bias assessments, and R code.

[DOCX File, 4775 KB - mental_v13i1e63088_app1.docx]

Checklist 1

PRISMA-P (Preferred Reporting Items for Systematic Review and Explanation Meta-Analysis Protocols) Checklist.

[PDF File, 94 KB - mental_v13i1e63088_app2.pdf]

References

1. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. Lancet Psychiatry 2016 Feb;3(2):171-178. [doi: 10.1016/S2215-0366(15)00505-2] [Medline: 26851330]
2. 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]
3. Bradburn NM, Rips LJ, Shevell SK. Answering autobiographical questions: the impact of memory and inference on surveys. Science 1987 Apr 10;236(4798):157-161. [doi: 10.1126/science.3563494] [Medline: 3563494]
4. Tversky A, Kahneman D. Availability: A heuristic for judging frequency and probability. Cogn Psychol 1973 Sep;5(2):207-232. [doi: 10.1016/0010-0285(73)90033-9]
5. Nelson EA, Maruish ME, Axler JL. Effects of discharge planning and compliance with outpatient appointments on readmission rates. Psychiatr Serv 2000 Jul;51(7):885-889. [doi: 10.1176/appi.ps.51.7.885] [Medline: 10875952]
6. Mitchell AJ, Selmes T. Why don't patients attend their appointments? Maintaining engagement with psychiatric services. Adv Psychiatr Treat 2007 Nov;13(6):423-434. [doi: 10.1192/apt.bp.106.003202]
7. Tiemens B, Kloos M, Spijker J, Ingenhoven T, Kampman M, Hendriks GJ. Lower versus higher frequency of sessions in starting outpatient mental health care and the risk of a chronic course; a naturalistic cohort study. BMC Psychiatry 2019 Jul 24;19(1):228. [doi: 10.1186/s12888-019-2214-4] [Medline: 31340791]
8. McQueenie R, Ellis DA, McConnachie A, Wilson P, Williamson AE. Morbidity, mortality and missed appointments in healthcare: a national retrospective data linkage study. BMC Med 2019 Jan 11;17(1):2. [doi: 10.1186/s12916-018-1234-0] [Medline: 30630493]
9. Mirah. URL: https://www.mirah.com [accessed 2025-12-18]
10. NeuroFlow. URL: https://www.neuroflow.com/ [accessed 2025-12-18]
11. Owl. URL: https://www.owl.health/ [accessed 2025-12-18]
12. PCOMS analysis web application. Better Outcomes Now. URL: https://betteroutcomesnow.com/ [accessed 2025-12-18]
13. Goldberg SB, Buck B, Raphaely S, Fortney JC. Measuring psychiatric symptoms remotely: a systematic review of remote measurement-based care. Curr Psychiatry Rep 2018 Aug 28;20(10):81. [doi: 10.1007/s11920-018-0958-z] [Medline: 30155749]
14. Fortney JC, Unützer J, Wrenn G, et al. A tipping point for measurement-based care. Psychiatr Serv 2017 Feb 1;68(2):179-188. [doi: 10.1176/appi.ps.201500439] [Medline: 27582237]
15. Schneider S, Stone AA. Ambulatory and diary methods can facilitate the measurement of patient-reported outcomes. Qual Life Res 2016 Mar;25(3):497-506. [doi: 10.1007/s11136-015-1054-z] [Medline: 26101141]
16. Runyan JD, Steinke EG. Virtues, ecological momentary assessment/intervention and smartphone technology. Front Psychol 2015;6:481. [doi: 10.3389/fpsyg.2015.00481] [Medline: 25999869]
17. Guo T, Xiang YT, Xiao L, et al. Measurement-based care versus standard care for major depression: a randomized controlled trial with blind raters. Am J Psychiatry 2015 Oct;172(10):1004-1013. [doi: 10.1176/appi.ajp.2015.14050652] [Medline: 26315978]
18. Simon GE, Ralston JD, Savarino J, Pabiniak C, Wentzel C, Operskalski BH. Randomized trial of depression follow-up care by online messaging. J Gen Intern Med 2011 Jul;26(7):698-704. [doi: 10.1007/s11606-011-1679-8] [Medline: 21384219]
19. Meglic M, Furlan M, Kuzmanic M, et al. Feasibility of an eHealth service to support collaborative depression care: results of a pilot study. J Med Internet Res 2010 Dec 19;12(5):e63. [doi: 10.2196/jmir.1510] [Medline: 21172765]
20. Miller SD, Duncan BL, Brown J, Sorrell R, Chalk MB. Using formal client feedback to improve retention and outcome: making ongoing, real-time assessment feasible. J Brief Ther 2006;5:5-22 [FREE Full text]
21. Dyer K, Hooke GR, Page AC. Effects of providing domain specific progress monitoring and feedback to therapists and patients on outcome. Psychother Res 2016;26(3):297-306. [doi: 10.1080/10503307.2014.983207] [Medline: 25506654]
22. Eisen SV, Dickey B, Sederer LI. A self-report symptom and problem rating scale to increase inpatients' involvement in treatment. Psychiatr Serv 2000 Mar;51(3):349-353. [doi: 10.1176/appi.ps.51.3.349] [Medline: 10686242]
23. Dowrick C, Leydon GM, McBride A, et al. Patients' and doctors' views on depression severity questionnaires incentivised in UK quality and outcomes framework: qualitative study. BMJ 2009 Mar 19;338(1):b663. [doi: 10.1136/bmj.b663] [Medline: 19299474]
24. Finn SE, Tonsager ME. Information-gathering and therapeutic models of assessment: complementary paradigms. Psychol Assess 1997;9(4):374-385. [doi: 10.1037//1040-3590.9.4.374]
25. Kendrick T, El-Gohary M, Stuart B, et al. Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults. Cochrane Database Syst Rev 2016 Jul 13;7(7):CD011119. [doi: 10.1002/14651858.CD011119.pub2] [Medline: 27409972]
26. Jensen-Doss A, Haimes EMB, Smith AM, et al. Monitoring treatment progress and providing feedback is viewed favorably but rarely used in practice. Adm Policy Ment Health 2018 Jan;45(1):48-61. [doi: 10.1007/s10488-016-0763-0] [Medline: 27631610]
27. Walsh AEL, Naughton G, Sharpe T, et al. A collaborative realist review of remote measurement technologies for depression in young people. Nat Hum Behav 2024 Mar;8(3):480-492. [doi: 10.1038/s41562-023-01793-5] [Medline: 38225410]

28. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097. [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
29. Machleid F, Michnevich T, Huang L, et al. Remote measurement based care (RMBC) interventions for mental health—protocol of a systematic review and meta-analysis. PLoS ONE 2024;19(2):e0297929. [doi: 10.1371/journal.pone.0297929] [Medline: 38363769]
30. da Costa Santos CM, de Mattos Pimenta CA, Nobre MRC. The PICO strategy for the research question construction and evidence search. Rev Lat Am Enfermagem 2007;15(3):508-511. [doi: 10.1590/s0104-11692007000300023] [Medline: 17653438]
31. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines: World Health Organization; 1992. URL: https://cdn.who.int/media/docs/default-source/classification/other-classifications/9241544228_eng.pdf [accessed 2025-12-18]
32. Diagnostic and Statistical Manual of Mental Disorders, 5th edition: American Psychiatric Association; 2013. [doi: 10.1176/appi.books.9780890425596]
33. RStudio: integrated development for R. RStudio. URL: http://www.rstudio.com/ [accessed 2025-12-18]
34. Viechtbauer W. Metafor: meta-analysis package for R. CRAN – The Comprehensive R Archive Network. URL: https://CRAN.R-project.org/package=metafor [accessed 2025-12-18]
35. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011 Feb 10;342:d549. [doi: 10.1136/bmj.d549] [Medline: 21310794]
36. Laursen SL, Helweg-Jørgensen S, Langergaard A, et al. Mobile diary app versus paper-based diary cards for patients with borderline personality disorder: economic evaluation. J Med Internet Res 2021 Nov 11;23(11):e28874. [doi: 10.2196/28874] [Medline: 34762057]
37. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration’s tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343:d5928. [doi: 10.1136/bmj.d5928] [Medline: 22008217]
38. Röver C. Bayesian random-effects meta-analysis using the bayesmeta R package. J Stat Softw 2020;93(6):1-51. [doi: 10.18637/jss.v093.i06]
39. Deeks JJ, Higgins JPT, Altman DG. Chapter 10: Analysing data and undertaking meta-analyses. In: Cochrane Handbook for Systematic Reviews of Interventions Version 62: Cochrane; 2021, Vol. 3. URL: https://training.cochrane.org/handbook/current/chapter-10 [accessed 2025-12-18]
40. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health - Care Evaluation, 1st edition: Wiley; 2003. [doi: 10.1002/0470092602]
41. Röver C, Friede T. Bayesmeta: bayesian random-effects meta-analysis and meta-regression. CRAN - The Comprehensive R Archive Network. 2025. URL: https://cran.r-project.org/web/packages/bayesmeta/index.html [accessed 2025-12-18]
42. Bürkner PC, Gabry J, Weber S, Johnson A, Modrak M, et al. Brms: bayesian regression models using “stan”. CRAN - The Comprehensive R Archive Network. 2025. URL: https://cran.r-project.org/web/packages/brms/index.html [accessed 2025-12-18]
43. Williams DR, Rast P, Bürkner PC. Bayesian meta-analysis with weakly informative prior distributions. PsyArXiv. Preprint posted online on Jan 11, 2018. [doi: 10.31234/osf.io/7tbrm]
44. Röver C, Bender R, Dias S, et al. On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. Res Synth Methods 2021 Jul;12(4):448-474. [doi: 10.1002/jrsm.1475] [Medline: 33486828]
45. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis: Chapman and Hall/CRC; 1995. [doi: 10.1201/9780429258411]
46. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
47. Chermahini MB, Eadie J, Agarwal A, et al. Comparing the efficacy of electronically delivered cognitive behavioral therapy (e-CBT) to weekly online mental health check-ins for generalized anxiety disorder-a randomized controlled trial [Comparaison de l’efficacité de la thérapie cognitivo-comportementale délivrée par voie électronique (e-TCC) aux contrôles hebdomadaires en ligne de santé mentale pour le trouble d’anxiété généralisée - un essai randomisé contrôlé]. Can J Psychiatry 2024 Sep;69(9):695-707. [doi: 10.1177/07067437241261933] [Medline: 39033431]
48. Cullen BA, Rodriguez K, Eaton WW, Mojtabai R, Von Mach T, Ybarra ML. Clinical outcomes from the texting for relapse prevention (T4RP) in schizophrenia and schizoaffective disorder study. Psychiatry Res 2020 Oct;292:113346. [doi: 10.1016/j.psychres.2020.113346] [Medline: 32750572]
49. Faurholt-Jepsen M, Frost M, Christensen EM, Bardram JE, Vinberg M, Kessing LV. The effect of smartphone-based monitoring on illness activity in bipolar disorder: the MONARCA II randomized controlled single-blinded trial. Psychol Med 2020 Apr;50(5):838-848. [doi: 10.1017/S0033291719000710] [Medline: 30944054]
50. Faurholt-Jepsen M, Lindbjerg Tønning M, Fros M, et al. Reducing the rate of psychiatric re-admissions in bipolar disorder using smartphones-The RADMIS trial. Acta Psychiatr Scand 2021 May;143(5):453-465. [doi: 10.1111/acps.13274] [Medline: 33354769]

51. Ebert D, Tarnowski T, Gollwitzer M, Sieland B, Berking M. A transdiagnostic internet-based maintenance treatment enhances the stability of outcome after inpatient cognitive behavioral therapy: a randomized controlled trial. Psychother Psychosom 2013;82(4):246-256. [doi: 10.1159/000345967] [Medline: 23736751]
52. Gallinat C, Moessner M, Apondo S, Thomann PA, Herpertz SC, Bauer S. Feasibility of an intervention delivered via mobile phone and internet to improve the continuity of care in schizophrenia: a randomized controlled pilot study. Int J Environ Res Public Health 2021 Nov 25;18(23):12391. [doi: 10.3390/ijerph182312391] [Medline: 34886117]
53. Lewis S, Ainsworth J, Sanders C, et al. Smartphone-enhanced symptom management in psychosis: open, randomized controlled trial. J Med Internet Res 2020 Aug 13;22(8):e17019. [doi: 10.2196/17019] [Medline: 32788150]
54. Spaniel F, Novak T, Bankovska Motlova L, et al. Psychiatrist's adherence: a new factor in relapse prevention of schizophrenia. A randomized controlled study on relapse control through telemedicine system. J Psychiatr Ment Health Nurs 2015 Dec;22(10):811-820. [doi: 10.1111/jpm.12251] [Medline: 26176646]
55. Freedman MJ, Lester KM, McNamara C, Milby JB, Schumacher JE. Cell phones for ecological momentary assessment with cocaine-addicted homeless patients in treatment. J Subst Abuse Treat 2006 Mar;30(2):105-111. [doi: 10.1016/j.jsat.2005.10.005] [Medline: 16490673]
56. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. Assessment 2023 Apr;30(3):825-846. [doi: 10.1177/10731911211067538] [Medline: 35016567]
57. Vachon H, Viechtbauer W, Rintala A, Myin-Germeys I. Compliance and retention with the experience sampling method over the continuum of severe mental disorders: meta-analysis and recommendations. J Med Internet Res 2019 Dec 6;21(12):e14475. [doi: 10.2196/14475] [Medline: 31808748]
58. Williams MT, Lewthwaite H, Fraysse F, Gajewska A, Ignatavicius J, Ferrar K. Compliance with mobile ecological momentary assessment of self-reported health-related behaviors and psychological constructs in adults: systematic review and meta-analysis. J Med Internet Res 2021 Mar 3;23(3):e17023. [doi: 10.2196/17023] [Medline: 33656451]
59. Burke LE, Shiffman S, Music E, et al. Ecological momentary assessment in behavioral research: addressing technological and human participant challenges. J Med Internet Res 2017 Mar 15;19(3):e77. [doi: 10.2196/jmir.7138] [Medline: 28298264]
60. Jones A, Remmerswaal D, Verveer I, et al. Compliance with ecological momentary assessment protocols in substance users: a meta-analysis. Addiction 2019 Apr;114(4):609-619. [doi: 10.1111/add.14503] [Medline: 30461120]
61. Cummings P. Meta-analysis based on standardized effects is unreliable. Arch Pediatr Adolesc Med 2004 Jun;158(6):595-597. [doi: 10.1001/archpedi.158.6.595] [Medline: 15184227]
62. Bauer M, Pfennig A, Schäfer M, Falkai P, editors. S3-Leitlinie Zur Diagnostik Und Therapie Bipolarer Störungen [Title in German]: Springer; 2020. [doi: 10.1007/978-3-662-61153-1]
63. Bipolar Disorder: Assessment and Management: National Institute for Health and Care Excellence (NICE); 2025. URL: https://www.nice.org.uk/guidance/cg185 [accessed 2025-12-18]
64. Hennemann S, Farnsteiner S, Sander L. Internet- and mobile-based aftercare and relapse prevention in mental disorders: a systematic review and recommendations for future research. Internet Interv 2018 Dec;14:1-17. [doi: 10.1016/j.invent.2018.09.001] [Medline: 30510909]
65. Rabung S, Harfst T, Kawski S, Koch U, Wittchen HU, Schulz H. Psychometrische Überprüfung einer verkürzten Version der »Hamburger Module zur Erfassung allgemeiner Aspekte psychosozialer Gesundheit für die therapeutische Praxis« (HEALTH-49) [Title in German]. Z Psychosom Med Psychother 2009 Apr 1;55(2):162-179. [doi: 10.13109/zptm.2009.55.2.162]
66. MacKrill K, Groom KM, Petrie KJ. The effect of symptom-tracking apps on symptom reporting. Br J Health Psychol 2020 Nov;25(4):1074-1085. [doi: 10.1111/bjhp.12459] [Medline: 32790051]
67. Eisner E, Faulkner S, Allan S, et al. Barriers and facilitators of user engagement with digital mental health interventions for people with psychosis or bipolar disorder: systematic review and best-fit framework synthesis. JMIR Ment Health 2025 Jan 20;12:e65246. [doi: 10.2196/65246] [Medline: 39832352]
68. Lupton D. The Quantified Self: Polity; 2016. URL: https://www.politybooks.com/bookdetail?book_slug=the-quantified-self--9781509500598 [accessed 2025-12-18]
69. Bless JJ, Hjelmervik H, Torsheim T, et al. Temporal signatures of auditory verbal hallucinations: an app-based experience sampling study. Schizophr Res 2020 Jan;215:442-444. [doi: 10.1016/j.schres.2019.11.020] [Medline: 31780342]
70. Barrio P, Ortega L, López H, Gual A. Self-management and shared decision-making in alcohol dependence via a mobile app: a pilot study. Int J Behav Med 2017 Oct;24(5):722-727. [doi: 10.1007/s12529-017-9643-6] [Medline: 28236288]
71. Gire N, Caton N, McKeown M, et al. 'Care co-ordinator in my pocket': a feasibility study of mobile assessment and therapy for psychosis (TechCare). BMJ Open 2021 Nov 16;11(11):e046755. [doi: 10.1136/bmjopen-2020-046755] [Medline: 34785541]
72. Liao Y, Skelton K, Dunton G, Bruening M. A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: an adapted STROBE Checklist for Reporting EMA Studies (CREMAS). J Med Internet Res 2016 Jun 21;18(6):e151. [doi: 10.2196/jmir.4954] [Medline: 27328833]
73. Heron KE, Everhart RS, McHale SM, Smyth JM. Using mobile-technology-based ecological momentary assessment (EMA) methods with youth: a systematic review and recommendations. J Pediatr Psychol 2017 Nov 1;42(10):1087-1107. [doi: 10.1093/jpepsy/jsx078] [Medline: 28475765]

74. Puhan MA, Soesilo I, Guyatt GH, Schünemann HJ. Combining scores from different patient reported outcome measures in meta-analyses: when is it justified? Health Qual Life Outcomes 2006 Dec 7;4:94. [doi: 10.1186/1477-7525-4-94] [Medline: 17156420]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
CONSORT-MHEALTH: Consolidated Standards of Reporting Trials-Mobile Health
***DSM*:** *Diagnostic and Statistical Manual of Mental Disorders*
EMA: ecological momentary assessment
***ICD*:** *International Classification of Diseases*
MBC: measurement-based care
MMH: mobile mental health
PICOS: population, intervention, comparison, outcome, and study
PRISMA-P: Preferred Reporting Items for Systematic Review and Explanation Meta-Analysis Protocols
RCT: randomized controlled trial
RMBC: remote measurement-based care
RoB-2: Cochrane risk-of-bias tool for randomized trials version 2
SMD: standardized mean difference

Edited by J Torous; submitted 10.Jun.2024; peer-reviewed by FZ Karimi, F Xu; accepted 21.Oct.2025; published 28.Jan.2026.

Please cite as:
Machleid F, Michnevich T, Huang L, Schröder-Frerkes L, Wiegmann C, Muffel T, Kaminski J
Remote Measurement-Based Care Interventions for Mental Health: Systematic Review and Meta-Analysis
JMIR Ment Health 2026;13:e63088
URL: https://mental.jmir.org/2026/1/e63088
doi:10.2196/63088

Digital Humans for Depression Assessment and Intervention Support: Scoping Review

Jiashuo Cao[1], MSc; Wujie Gao[2], MA; Ruoyu Wen[3], MA; Chen Li[2,4], PhD; Simon Hoermann[3], PhD; Nilufar Baghaei[5], PhD; Mark Billinghurst[1], PhD

[1]Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand
[2]Department of Applied Social Sciences, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)
[3]School of Product Design, University of Canterbury, Christchurch, New Zealand
[4]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)
[5]School of Electrical Engineering and Computer Science, The University of Queensland, Brisbane, Australia

Corresponding Author:
Jiashuo Cao, MSc
Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand

Abstract

Background: The growing global burden of mental health disorders has intensified the search for scalable, accessible, and cost-effective interventions. Conversational agents in the form of digital humans have emerged as promising tools to deliver mental health support across diverse populations and settings.

Objective: This scoping review aimed to analyze the role of digital humans in depression management, identifying their specific applications in both diagnostic processes and therapeutic interventions. Additionally, it aimed to evaluate the design choices implemented in digital human systems, including their appearance, interaction modalities, back-end intelligence systems, and the various roles they assume.

Methods: Following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines, we systematically searched peer-reviewed literature across major databases, including ACM Digital Library, IEEE Xplore, Web of Science, and PubMed, to capture both psychological and technological perspectives. The search query included a wide variety of synonyms for digital humans and depression: ("avatar" OR "virtual agent" OR "embodied conversational agent" OR "relational agent" OR "digital human" OR "virtual human" OR "virtual character") AND ("Major Depressive Disorder" OR "Depression"). Studies were included if they described the development, implementation, or evaluation of digital humans designed to support mental health outcomes. Data were charted on agent design, therapeutic approach, target population, delivery context, and reported effectiveness.

Results: In total, 20 studies (2010 - 2024) were included. Depression assessment studies comprised 35% (n=7), interventions 55% (n=11), and combined approaches 10% (n=2). Assessment protocols included the questionnaires Patient Health Questionnaire-9 and Very Short Visual Analog Scale of the Center for Epidemiologic Studies Depression Scale - Visual Analog Scale - Very Short version, semistructured interviews based on *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* criteria, and interactive tasks designed to elicit emotional responses. Intervention approaches used cognitive behavioral therapy, psychoeducation, compassion-focused therapy, and avatar therapy. Digital humans assumed 5 distinct roles: interviewer (n=6), facilitator (n=3), counselor (n=3), educator (n=3), and actor (n=5). Interviewers primarily appeared in assessment studies, presenting structured questions. Counselors engaged in therapeutic dialogues, while educators delivered psychoeducational content. Facilitators assisted participants in achieving system goals. Actors portrayed specific emotions or dysfunctional beliefs to facilitate therapeutic processes. Studies highlighted digital humans' utility in enhancing diagnostic processes and therapeutic interventions, noting the potential for transformation through physiological data integration.

Conclusions: This study demonstrates that digital humans represent a transformative advancement in depression management, offering innovative applications across both assessment and intervention phases. The evidence reveals digital humans' effectiveness in replicating traditional therapeutic roles while providing unique advantages, including 24/7 accessibility, reduced stigma, consistent care delivery, and personalized support. Digital humans can successfully function to establish therapeutic alliances and elicit meaningful engagement comparable with human providers. Findings underscore the need for continued research to fully realize digital humans' potential in addressing depression-specific needs, advocating for expansion into diverse therapeutic scenarios, and exploration of unexplored digital human applications.

(JMIR Ment Health 2026;13:e79954) doi:10.2196/79954

KEYWORDS
digital human; virtual agent; mental health; depression; embodied conversational agent

Introduction

Depression is a prevalent mental health disorder affecting millions of individuals worldwide, with the World Health Organization identifying it as a leading cause of disability globally [1]. Characterized by persistent sadness, lack of interest in daily activities, and a multitude of physical and psychological symptoms, depression can severely impact an individual's quality of life [2]. Psychotherapy's effectiveness in treating depression across diverse populations and settings is widely recognized [3,4]. It has been demonstrated to be as effective as antidepressant medication for individuals with mild to moderate depression and is often the preferred initial treatment option for patients [5,6]. Furthermore, psychotherapy may surpass pharmacological treatment of depression in terms of long-term effectiveness [7].

In recent years, technology has been integrated into mental health therapy, allowing for novel ways of therapy as well as reaching people who otherwise would not have access to therapy [8]. One of those recent technology advancements is embodied conversational agents, often in the form of virtual human characters—in this paper referred to as digital humans [9]—which have shown potential for health assessments and interventions [10].

Hence, the emergence of virtual characters as a supportive tool for depression highlights a significant trend, which is propelled by advancements in artificial intelligence and computer graphics [11]. These digital characters provide increasingly lifelike, responsive, and immersive interactions, capable of perceiving and reacting to the emotional states of users [9]. They offer customized support and interventions, paralleling the capabilities of human therapists to a notable extent [12]. This technological progression enables the delivery of online therapy exercises, mindfulness techniques, and emotional support accessible for a wide range of people. Individuals can engage with these therapeutic resources from their own homes, effectively overcoming obstacles such as societal stigma, geographic barriers, and the prohibitive costs often associated with traditional therapeutic services [13,14]. Moreover, the incorporation of machine learning algorithms allows these virtual agents to evolve through user interaction, enhancing their supportive capabilities over time and furnishing a personalized therapeutic experience [15]. As these technological innovations advance, virtual characters are poised to become a fundamental component of mental health care. Based on current literature, while there have been reviews and meta-analyses examining the use of digital humans and digital interventions in health contexts, a critical gap exists in understanding how digital humans have been specifically used to support depression and what detailed design choices have been made to adapt them for this purpose. For instance, Ma et al [11] conducted a meta-analysis of virtual human interventions across various health conditions, providing valuable insights into intervention outcomes, but did not focus specifically on digital humans in depression contexts nor examine the design characteristics that enable therapeutic interactions. Chattopadhyay et al [12] explored the application of virtual humans in health care systems broadly, emphasizing implementation contexts and user perceptions, but did not analyze the technical and aesthetic design decisions—such as appearance choices, interaction modalities, or behavioral capabilities—that shape these systems. Moshe et al [14] discussed the effectiveness of digital interventions for depression but focused primarily on app-based and online platforms, lacking in-depth exploration of embodied conversational agents and their unique design considerations. Thus, a comprehensive analysis of digital humans' roles in depression management, revealing their exact value, is needed. To fulfill the goal of unpacking the benefits of using digital humans to support depression, we investigate two research questions (RQs):

- First, how are digital humans used in the assessment and intervention of depression?
- Second, what design considerations were made to adapt digital humans for depression assessment and intervention?

This study presents 2 primary contributions. First, it provides a detailed overview of the current state of research regarding the use of digital humans in supporting individuals with depression. This encompasses a thorough analysis of the various types of support services available and the specific design choices regarding the implementation of digital humans. Second, it identifies and proposes several areas for future research within this domain that merit further investigation.

Methods

Overview

We conducted the scoping review in February 2025. Scoping reviews aim to facilitate the formulation of pertinent RQs by rapidly identifying and categorizing existing evidence within a given field [16]. Our methodology was anchored in the guidelines set forth by Munn et al [17], complemented by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist's extension, specifically the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) framework [18], which is designed for scoping reviews.

Definitions

Digital Human

In this review, the definition of digital human is equivalent to the definition of virtual human. As described by Traum [19], a virtual human is an "artificial agent that includes both a visual body with a humanlike appearance and range of observable behaviors, and a cognitive component that can make decisions and control the behaviors to engage in human-like activities." Although this definition is comprehensive, there are still some ambiguities, such as the judgment of humanlike appearance and humanlike activities. To provide a clearer set of criteria to assist us in filtering articles, we have defined a digital human as an agent with 3 criteria inspired by the 10 traits suggested by Burden and Savin-Baden [9]. A digital human (1) visually possesses realistic appearance characteristics of a human, including facial features and body proportions; (2) is capable of performing nonverbal behaviors, including body movements and facial expressions; and (3) must engage in bidirectional communication, understanding, and responding to verbal and nonverbal cues from users.

If the virtual collocutor described in the reviewed article meets all of the criteria, then we consider it to be a digital human.

Support for Depression

Support for depression encompasses a holistic approach that integrates both direct and indirect methods to aid individuals in managing and overcoming the condition. This comprehensive support system is essential for addressing the multifaceted nature of depression, which affects individuals emotionally, physically, and socially [3]. Following the framework established by Cuijpers [20], we distinguish between direct and indirect support based on their primary target. Indirect support focuses on problems related to depression—such as social isolation, lifestyle factors, or caregiver burden—where interventions address associated factors rather than depression as the primary clinical target. In contrast, direct support explicitly targets depression assessment (eg, administering Patient Health Questionnaire-9 [PHQ-9], conducting diagnostic interviews based on *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* [*DSM-5*] criteria) or depression-specific therapeutic interventions (eg, cognitive behavioral therapy [CBT] for negative thought patterns and compassion-focused therapy [CFT] for self-criticism). In this paper, we narrow the scope to direct support provided to individuals with depression, as this aligns with our RQs that specifically examine how digital humans function within depression assessment and intervention protocols. This means that the primary contribution of selected papers is to provide assistance to individuals undergoing assessment for depression and receiving treatment for it.

Information Sources and Search Strategy

Given the interdisciplinary nature of the topic, the search was carried out across four distinct digital libraries, spanning both psychological and technological fields: ACM Digital Library, IEEE Xplore, Web of Science, and PubMed. Based on the RQs and definition of digital human, we constructed the search query "("avatar" OR "virtual agent" OR "embodied conversational agent" OR "relational agent" OR "digital human" OR "virtual human" OR "virtual character") AND ("Major Depressive Disorder" OR "Depression")." We did not include context-related terms in the search query (such as assessment, screening, and intervention) because we aim to find as many records as possible of digital humans supporting depression. Later, we will filter out the articles that meet the requirements of this paper using exclusion criteria.

Eligibility Criteria

Based on our definition of digital human and RQs, we created some inclusion and exclusion criteria (Textbox 1). The exclusion criteria were used to remove papers from consideration.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Full text available.
- Paper must be published in English.
- Paper provides details of digital human.
- At least one empirical study has been conducted on depression.
- The purpose of the study is to support depression assessment or intervention.
- The study included interaction between participants and digital humans.

Exclusion criteria

The measured outcome is not related to depression or major depressive disorder.

- The contribution of work does not support depression assessment or intervention.
- No digital human described in the paper.
- No interaction between participant and digital human.

Search Result and Study Selection

The search led to the identification of 1031 publications. After excluding duplicate records, we amassed a total of 909 papers. The assessment process for these papers was conducted in 2 distinct phases. Initially, a preliminary assessment based on titles and abstracts was carried out, with the first 50 papers being collaboratively reviewed by the first, second, and third authors (JC, WG, and RW). Following a consensus on the inclusion and exclusion decisions regarding these 50 papers, the remaining papers were evenly distributed among the same 3 authors for review. This first phase resulted in the identification of 52 pertinent papers. Subsequently, 3 additional publications were manually sourced, culminating in a total of 55 new references. These references were then apportioned among 4 authors (JC, WG, RW, and CL) for an in-depth assessment of the full texts, guided by the predefined inclusion and exclusion criteria. This process led to the selection of 20 papers [21-40] for subsequent data extraction.

Data Charting Process and Data Items

Data extraction was carried out by the first and second authors (JC and WG), ensuring a thorough and collaborative approach to gathering information. This process involved detailing the characteristics of each study. To facilitate coordination and accuracy, the extracted data were compiled into a shared Google Sheet. To further enhance the reliability of the data collection process, the first author (JC) conducted a comprehensive review of all data extractions, ensuring consistency and accuracy across the dataset. Specifically, the following data items were extracted from selected papers:

- General characteristics: title, authors, year of publication, journal or proceedings, and study aims (Table 1).
- Study design and findings: setting, sample size, protocol, role of digital human, and findings (Table 2).
- Digital human design: appearance, display device, system type, and input and output modality (Table 2).

Table . Summary of selected studies.

Study	Journal or proceedings	Experiment settings	Participants	Service	Aim of study
Jaiswal et al [21]	19th ACM[a] International Conference on Intelligent Virtual Agents	Laboratory	n=55	Assessment	Investigated digital human administered questionnaires for depression and anxiety
Egede et al [22]	21st ACM International Conference on Intelligent Virtual Agents	Laboratory	n=56	Assessment	Evaluated the effectiveness of digital human-mediated tasks in depression assessment
Wolters et al [23]	HCI[b] KOREA 2015	Field	n=4	Assessment and intervention	Explored the use of personal monitoring system with digital human integrated
Baghaei et al [24]	2021 ACM SIGCHI[c] Symposium on Engineering Interactive Computing Systems	Laboratory	n=23	Intervention	Investigated the feasibility of CFT[d] with digital human in VR[e]
Luerssen and Hawke [25]	18th International Conference on Intelligent Virtual Agents	Field	n=9	Intervention	Evaluated the effectiveness of LiCBT[f] conducted by virtual coach
Takemoto et al [26]	Journal of Eye Movement Research	Laboratory	n=27	Assessment	Explored the possibility of using digital human communication and eye tracking to detect depression
Bresó et al [27]	Expert Systems	Laboratory	n=60	Intervention	Evaluated the usability and acceptability of a digital human that could identify and provide an early intervention for depression
Takemoto et al [28]	Frontiers in Digital Health	Laboratory	n=27	Assessment	Explored the possibility of using digital human communication and facial expression to detect depression
DeVault et al [29]	14th annual SIGdial[g] Meeting on Discourse and Dialog	Laboratory	n=43	Assessment	Explored the presence of indicators of psychological distress in semistructured digital human interview
Ashrafi et al [30]	International Conference on Human-Computer Interaction 2024	Laboratory	n=22	Intervention	Explored the effect of visual similarity of digital human in psychotherapy
Wu et al [31]	IEEE[h] Transactions on Affective Computing	Laboratory	n=168	Assessment	Evaluated the accuracy of using digital human for automatic depression-level stratification on mobile devices
Kocur et al [32]	Frontiers in Psychiatry	Clinic	n=54	Intervention	Evaluated the effectiveness of a computer-assisted, avatar-based therapy in reducing dysfunctional beliefs in depressive inpatients

Study	Journal or proceedings	Experiment settings	Participants	Service	Aim of study
Burton et al [33]	Telemedicine and Telecare	Field	n=28	Assessment and intervention	Explored the use of personal monitoring system with digital human integrated
Shamekhi et al [34]	16th International Conference on Intelligent Virtual Agents	Field	n=20	Intervention	Investigated the use of digital human to review material after medical visit
Halim et al [35]	JMIR[i] Mental Health	Laboratory	n=36	Intervention	Evaluated the usability and acceptability of CFT with digital human in VR
Bickmore et al [36]	Interacting with Computers	Clinic	n=131	Intervention	Explored the use of digital human explain self-care regimen to patients with depressive symptoms
Ring et al [37]	2016 ACM CHI[j] workshop on Computing and Mental Health	Laboratory	n=10	Intervention	Investigated the efficacy of virtual therapist for depression counseling
Shamekhi et al [38]	2017 International Conference on Persuasive Technology	Field	n=154	Intervention	Evaluated the effectiveness of using digital humans for stress management
Philip et al [39]	Scientific Reports	Clinic	n=179	Assessment	Evaluated the performance and acceptability of using digital human as a diagnostic tool for depression through interview
Hidding et al [40]	Behaviour Research and Therapy	Laboratory	n=68	Intervention	Investigated the effects of a VR intervention on self-criticism and self-compassion, including the use of a digital human

[a]ACM: Association for Computing Machinery.
[b]HCI: Human-Computer Interaction.
[c]SIGCHI: Special Interest Group on Computer-Human Interaction.
[d]CFT: compassion-focused therapy.
[e]VR: virtual reality.
[f]LiCBT: low-intensity cognitive behavioral therapy.
[g]SIGdial: Special Interest Group on Discourse and Dialogue.
[h]IEEE: Institute of Electrical and Electronics Engineers.
[i]JMIR: Journal of Medical Internet Research.
[j]CHI: Conference on Human Factors in Computing Systems.

Table . Details of digital human in selected studies.

Study	Protocol	Role of digital human	Appearance	Input modality	Output modality	Back-end intelligence
Jaiswal et al [21]	Interview	Interviewer	Full body	Speech	Text, speech, behavior	Rule-based
Egede et al [22]	Computer-based interactive task	Facilitator	Full body	Speech	Text, speech, behavior	Scripted
Wolters et al [23]	PHQ-9[a], CBT[b]	Facilitator	Upper body	Text and speech	Speech, behavior	Scripted
Baghaei et al [24]	CFT[c]	Actor	Full body	Speech	Speech, behavior	Scripted
Luerssen and Hawke [25]	LiCBT[d]	Counselor	Upper body	Text and speech	Text, speech, behavior	Rule-based
Takemoto et al [26]	Interview	Interviewer	Upper body	Speech	Speech, behavior	Scripted
Bresó et al [27]	CBT	Counselor	Head-only	Speech	Speech, behavior	Rule-based
Takemoto et al [28]	Interview	Interviewer	Upper body	Speech	Speech, behavior	Scripted
DeVault et al [29]	Interview	Interviewer	Full body	Speech	Speech, behavior	Wizard of Oz
Kocur et al [32]	Avatar Therapy	Actor	Head-only	Speech	Speech, behavior	Wizard of Oz
Burton et al [33]	PHQ-9, CBT	Facilitator	Upper body	Text and speech	Speech, behavior	Scripted
Shamekhi et al [34]	Psychoeducation	Educator	Upper body	Touch	Text, speech, behavior	Scripted
Halim et al [35]	CFT	Actor	Full body	Speech	Speech, behavior	Recorded playback
Halim et al [36]	Psychoeducation	Educator	Full body	Touch	Text, speech, behavior	Scripted
Ring et al [37]	CBT	Counselor	Upper body	Text, speech, and video	Text, speech, behavior	Rule-based
Shamekhi et al [38]	Psychoeducation	Educator	Upper body	Touch	Text, speech, behavior	Scripted
Philip et al [39]	Interview	Interviewer	Upper body	Speech	Speech, behavior	Scripted
Ashrafi et al [30]	Avatar Therapy	Actor	Upper body	Speech	Speech, behavior	Scripted
Hidding et al [40]	CBT	Actor	Full body	Speech	Speech, behavior	Wizard of Oz
Wu et al [31]	Interview	Interviewer	Upper body	Speech	Speech, behavior	Scripted

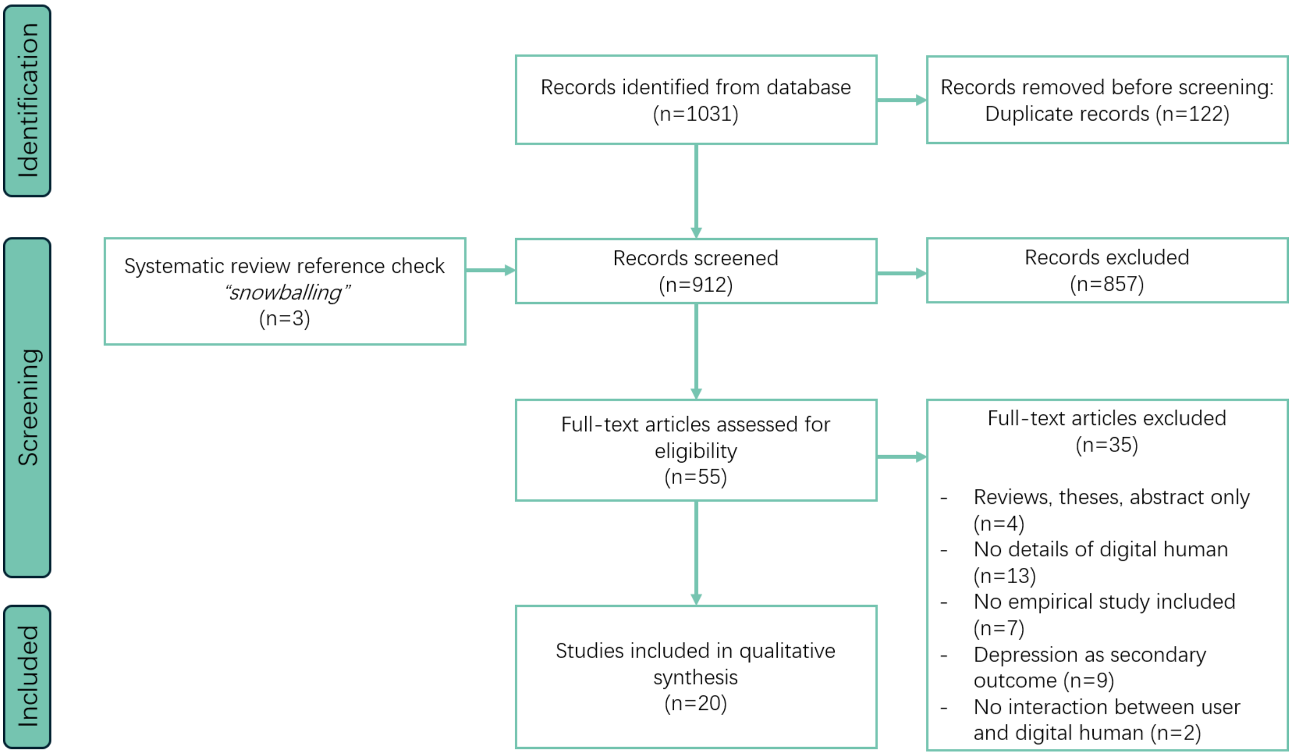
[a]PHQ-9: Patient Health Questionnaire-9.
[b]CBT: cognitive behavioral therapy.
[c]CFT: compassion-focused therapy.
[d]LiCBT: low-intensity cognitive behavioral therapy.

Results

Summary of Selected Literature

We completed the literature search in February 2025, and after the assessment stage, a total of 20 articles [21-40] were selected in this review, as detailed in Table 1. The detailed inclusion and exclusion criteria for studies at each phase of our review are visually represented in the PRISMA flow diagram (Figure 1).

Figure 1. Literature screening and selection diagram following PRISMA guidelines. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.



The selected papers were published between 2010 and 2024, with those from the past 4 years accounting for 45% of the total (n=9) [22,24,26,28,30-32,35,40]. Of these, 10 papers (50%) [26-28,31-33,35,36,39,40] were presented in peer-reviewed journals, while the remaining articles were published in conference proceedings (n=10, 50%) [21-25,29,30,34,37,38].

More than half of the studies were conducted in research laboratory settings (n=12, 60%) [21,22,24,26-31,35,37,40], 5 experiments (25%) [23,25,33,34,38] took place in the participants' everyday environments, and 3 (15%) [32,36,39] were conducted in clinical settings. Regarding participant numbers, 11 studies (55%) [23-26,28-30,33-35,37] involved fewer than 50 participants, 5 studies (25%) [21,22,27,32,40] had between 50 and 100 participants, and 4 studies (20%) [31,36,38,39] included more than 100 participants. Of these 4 studies with over 100 participants, 2 [36,39] were conducted in clinical environments, 1 [38] took place in an everyday environment, and 1 [31] in a laboratory setting.

In all the studies, those aimed at assessing depression accounted for 35% (n=7) [21,22,26,28,29,31,39], while studies focused on interventions constituted 55% (n=11) [24,25,27,30,32,34-38,40]. The remaining 2 studies [23,33], accounting for 10%, included both assessment and intervention.

Summary of Support Services With Digital Human

We analyzed the selected studies primarily focusing on 3 factors: the protocols used, the roles played by digital humans within the service, and the experimental findings. Regarding protocols, commonly used assessment tools include questionnaires and interviews, as well as specially designed computer-based interactive tasks. Additionally, therapeutic approaches, such as psychoeducation, CBT [41], and CFT [42], have also been used.

The roles of digital humans were primarily as interviewers, facilitators, and educators, actively interacting with participants. Notably, in the study by Baghaei et al [24], the digital human was designed as an actor, requiring proactive interaction from the participants. The findings from these experiments focused on 2 main areas—the usability and acceptability of digital humans in health care applications, and the potential for integrating digital humans into various protocols.

Applied Protocol

Assessment

The applied protocol in assessment can be mainly divided into 3 categories—questionnaire, interview, and computer-based interactive task.

The first category is questionnaires, accounting for 3 of the studies [21,23,33]. In this category, the digital human will ask users the questionnaire questions one by one and collect their responses. Only answers that correspond to existing options will be collected; otherwise, the digital human will continue asking the user until obtaining a usable answer. The questionnaires that were used include PHQ-9 [43] and Center for Epidemiologic Studies Depression Scale - Visual Analog Scale - Very Short version (CES-D-VAS-VS) [44]. PHQ-9 is a brief self-report tool that consists of 9 items, each of which is scored on a scale from 0 (not at all) to 3 (nearly every day), which are directly derived from the diagnostic criteria for major depressive disorder in the *DSM*. The CES-D-VAS-VS is an adaptation of the traditional Center for Epidemiologic Studies - Depression Scale, which is a 20-item questionnaire used to measure depressive symptoms in the general population. The CES-D-VAS-VS incorporates a visual analog scale, enhancing its sensitivity by allowing patients to mark their symptom severity along a continuum.

The second category is assessments through interviews, accounting for 5 of the studies [26,28,29,31,39]. As an essential tool in depression assessment, interviews are commonly used in clinical settings where a nuanced understanding of the patient's condition is crucial [45]. In the 5 studies that used interviews for depression assessment, 2 [26,28] focused on interviewing about negative topics, such as war and loneliness. These topics were chosen because they significantly impact vocal, visual, and verbal features, which are critical in detecting depression. The study by DeVault et al [29] based its interview content on observations from face-to-face interviews conducted in a clinical setting. This approach aimed to mimic real-life interactions and assess how patients express their symptoms naturally. Wu et al [31], through collaboration with clinicians, designed 93 questions about depression, anxiety, hypomania, and family relationships. This series of questions was specifically designed to vary depending on positive, negative, and neutral emotions. The study by Philip et al [39] structured its interview around the *DSM-5* [46] criteria for major depressive disorder. This method ensured that the interviews were comprehensive and aligned with established diagnostic standards, facilitating a more systematic approach to identifying depressive symptoms based on the latest psychiatric guidelines.

A computer-based interactive task was only used in 1 study [22]. In this study, researchers and mental health experts collaboratively designed 4 different types of tasks—mimicking, dyadic interaction, digital treatment, and psychometric. The design of these tasks aimed to be sufficiently engaging to prompt use without the need for physician guidance, executable on mobile devices, and effective at eliciting behavioral traits relevant to digitally assess mental health.

Intervention

There are a total of 13 (65%) studies focusing on the intervention phase. The therapeutic methods involved include CBT [23,25,27,33,37,40,47], psychoeducation [34,36,38], CFT [24,35], and avatar therapy [30,32].

CBT [48], as one of the most common therapy approaches in studies, aimed at teaching individuals how to recognize negative patterns of thought, evaluate their validity, and replace them with healthier ways of thinking. Among all the studies that applied CBT, they mainly focused on cognitive restructuring [49] and behavioral activation techniques [50]. In cognitive restructuring, participants need to recognize negative thoughts related to a recent event and are then encouraged to contemplate a more positive interpretation, and behavioral activation is a CBT technique that motivates participants to undertake avoided activities and engage in tasks that bring pleasure and accomplishment. Luerssen and Hawke [25] applied low-intensity CBT [51] in their study. Compared with traditional CBT, low-intensity CBT involves simpler interventions that can be delivered by practitioners who are not necessarily clinical psychologists and is more standardized, using general strategies that apply broadly to everyone with similar symptoms, while CBT is highly personalized and involves developing a specific therapeutic strategy for each patient's unique problems and needs.

Psychoeducation is a fundamental component of depression intervention that involves educating individuals about depression as a disorder, including its symptoms, causes, and treatment options [52]. In the studies by Shamekhi et al [34,38], the content of psychoeducation primarily focuses on the management of nutrition, physical activity, pain, stress, sleep, and depression and includes guiding patients through practice sessions, such as meditation and yoga. Meanwhile, in the study by Bickmore et al [36], the educational content focused on the postdischarge self-care regimen, including medications, follow-up appointments, exercise and diet regimens, and pending laboratory tests.

In the remaining 4 studies, 2 [24,35] used CFT [42], a psychological approach designed to promote mental and emotional healing by encouraging individuals to develop compassion for themselves and others. The other 2 studies [30,32] used avatar therapy, where therapists interact with clients using a computer-generated avatar. Ashrafi et al [30] conducted a computer-assisted avatar-based treatment for dysfunctional beliefs (CAT-DB), which uses an avatar to help patients engage in dialogue with their dysfunctional beliefs and confront them.

Role of Digital Human

In the selected studies, digital humans assumed 5 different roles—interviewer [21,26,28,29,31,39], facilitator [22,23,33], counselor [25,27,37], educator [34,36,38], and actor [24,30,32,35,40]. The interviewer primarily appears in assessment-type studies, where their main task is to present predesigned questions to participants and await their responses. Counselors also engage in conversation with participants, but their dialogues focus more on therapeutic interactions. For example, in the study by Ring et al [37], the digital human serves as a counselor, providing the first counseling session in a CBT intervention. Educators, designated specifically for psychoeducation, are responsible for conveying specific learning material to participants through visual and verbal means and assisting them in reviewing previously learned content. Facilitators focus on assisting participants in achieving specific goals requested by the system by providing task instructions and confirming task completion. Actor is a unique role compared with others; digital humans play specific characters to facilitate psychotherapy. In the study by Baghaei et al [24], the digital human acts in a state of negative emotion (angry, crying, etc) to elicit compassion from participants. In another study by Kocur et al [32], the digital human represents a human entity that continually expresses dysfunctional beliefs to participants, such as "You have to be perfect."

Feasibility and Effectiveness of Digital Human

Assessment

In studies applying digital humans to assessment, digital humans have been proven to be as effective in conducting interviews as showing real human videos [26,28] and filling out self- report questionnaires [21]. In the studies by DeVault et al [29] and Philip et al [39], data collected after interviews with digital humans were analyzed, and they concluded that these data have the potential to support automatic depression assessment based on dialogue systems involving digital humans and humans. Furthermore, in the experiment by Egede et al [22], it was found that participants exhibited greater body movements and more intense facial expressions under the guidance of a digital human (compared with text-only mode), which also supports the feasibility of using digital humans for automatic depression assessment.

Intervention

Many experimental results that apply digital humans in interventions have highlighted the potential for digital humans to establish positive relationships with participants. This includes factors such as positive responses to participants' emotional expressions, customizable appearances, nonjudgmental expressions, and good patience. This personal relationship can also lead to better acceptance and attitudes toward digital humans from participants. Several articles reported that digital humans made a positive impression on participants. In the experiment by Bickmore et al [36], 76% of participants reported that their experience talking with a digital human was better than their interactions with health providers, and they were more inclined to choose digital humans as their channel for receiving psychoeducation in the future. Additionally, several articles indicated a positive impact of digital humans on therapeutic outcomes. For example, in the experiment by Kocur et al [32], they compared the results of participants receiving treatment as usual (TAU) with those receiving CAT-DB+TAU and found significant differences in the reduction of depressive symptoms in the group who received CAT-DB+TAU.

Design of Digital Human

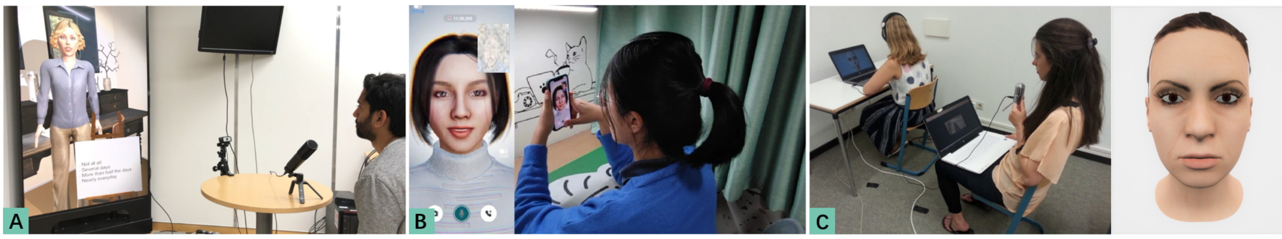
Overview

The design of digital humans across various studies showcases a diverse range of models, display devices, input or output modalities, and back-end intelligence, reflecting the adaptability of virtual characters to different applications. Based on the description of digital humans in the studies, we identified that the digital human designs in 3 studies [23,26,34] appeared repeatedly. We chose to retain the literature with more detailed descriptions of the digital human designs. After eliminating these duplicates, 17 unique digital human designs were included in the results.

Digital Human Presentation

In terms of body visibility, there are 7 full-body, 8 upper-body (head to chest), and 2 head-only digital humans (Figure 2). After analyzing the use cases for each digital human, we found that on mobile devices (such as phones and tablets), half-body is a more common setting [31,37], as this aligns better with how we typically see human bodies on these devices. The choice of full body is usually driven by device or scenario requirements. For example, digital humans in virtual reality (VR) all use full-body representations [24,35,40], and in the study by Stratou et al [53], they displayed a full-body digital human in larger screens to simulate the feeling of conversing with real people in reality. As mentioned by Bresó et al [27], they chose head-only because focusing on facial expressions makes it easier to convey different emotions and gives the digital human a higher degree of realism, which is a key issue in their study design. Regarding style design, the appearance of 2 digital humans leaned more toward a cartoon-like style [25,28]. Luerssen and Hawke [25] mentioned that their choice of cartoon-like style was to avoid the uncanny valley effect, while the rest of the digital humans adopted a more realistic style.

Figure 2. Example of digital human: (A) full body [21], (B) upper body [31], and (C) head-only [32].



Input and Output Modality

The majority of digital humans (n=15, 88.2%) support natural language input, with 2 studies [36,38] relying on interface touch input. Correspondingly, all digital humans support speech synthesis and audio output, but only 5 digital humans are capable of delivering empathy narration [22,27,29,30,35]. All digital humans exhibit lip movements while speaking, along with other bodily actions, such as blinking [28] and nodding [29]. Due to their presence in a 3D environment, the digital humans in the studies by Baghaei et al [24], Halim et al [35], and Wu et al [31] are capable of a wider range of actions, including walking, crying, showing anger, and more.

Back-End Intelligence

This scoping review identifies 4 types of back-end intelligence used in the development of digital humans—Scripted Systems, Rule-Based Systems, Wizard of Oz, and Recorded Playback. Each approach offers unique strengths and applications, contributing to the diverse landscape of digital human technology.

Scripted systems operate based on predetermined scripts that define the digital human's behavior and responses in a linear, predictable manner. This method is highly effective in scenarios where interactions are straightforward and the range of possible user inputs is limited. Among the 17 digital humans, more than half (n=10, 58.8%) used a scripted back end to control their behavior. For the digital human who used a scripted back end, the scenarios they face have standard dialogue processes, such as interviews with predefined topics [26,31], and tasks with specific instructions [22].

Rule-based systems are characterized by a set of predefined rules that guide the digital human's decision-making processes. These systems offer greater flexibility compared with scripted systems, as they can dynamically adapt to a wider range of user inputs through conditional logic. Furthermore, 4 digital humans use a rule-based dialogue system, with 3 digital humans acting as counselors in the studies [25,27,37], and 1 as an interviewer [21].

The Wizard of Oz [54] technique involves a human operator who controls the digital human's actions and responses in real-time, unbeknownst to the user. This method is commonly used in experimental and prototyping phases, allowing researchers to simulate the capabilities of a fully autonomous digital human before the underlying technology is fully developed. The digital humans in the experiments conducted by DeVault et al [29], Kocur et al [32], and Hidding et al [40] are directly controlled by professionals (certified counselors) to ensure participant safety and provide an experience closer to interacting with a real person.

Notably, in the experiment by Halim et al [35], the behavior of the digital human is derived from the participants' actions, a method called Recorded Playback. In this experiment, participants are required to show compassion to the digital human in the first stage, and the language and actions of the participants are recorded by the system. In the second stage, these recorded behaviors are performed by the digital human.

Discussion

Principal Findings

RQ 1: Digital Human Replication of Roles in Supporting Services

The aim of this scoping review was to identify the existing evidence on using digital humans to assist in assessment for depression and delivering interventions, and to unfold the design choices of these digital humans. We have found that many studies designed digital humans' purposes to be similar to certain roles in existing services. For example, a counselor conducting assessment through interviews [31], a therapist providing CBT session [27], and a nurse giving necessary information to patients [36]. The results of these studies can be considered preliminary affirmations of the effectiveness of digital humans in these roles.

Advantages of digital humans: Compared with real humans, digital humans have significant advantages in accessibility and availability, as they are essentially software programs. This means they can operate around the clock without the constraints of human limitations such as fatigue, working hours, or geographic location. This 24×7 availability ensures that support can be provided to users whenever they need it, which is particularly beneficial in emergency situations or for individuals in different time zones. Additionally, applying digital humans in the field of psychotherapy has a special advantage—the low social stigma caused by interaction. In selected studies, 2 articles [21,36] mentioned that in experimental interviews comparing interactions with digital humans and health providers (therapists or nurses), several participants felt more relaxed and safe communicating with the digital human. This advantage has been proven by the research by Lucas et al [55], which demonstrated that participants were more willing to disclose emotionally sensitive information, including expressing sadness more intensely, when they were interviewed by a digital human and believed the interviewer was a computer rather than a human. Another work by Loveys et al [56] indicates that the emotional expressiveness of digital humans, such as using an emotional voice, can enhance participants' comfort and emotional responses during interactions. These studies suggest that digital humans can create an environment where individuals feel more comfortable and less anxious when disclosing. Furthermore, the consistency and unbiased nature of digital humans ensure that all users receive the same level of care and attention, free from the potential biases or variability that can come with human providers. This standardization can be particularly important in ensuring equitable access to high-quality support services.

Designed for real-life scenarios: In the design of these digital humans, a key design choice has been to ensure that the digital human's appearance and behavior align with the local cultural context. For example, Takemoto et al [26] specifically mentioned that they designed the digital human's appearance based on the local population's characteristics. In the study by Wolters et al [23], besides the appearance, the digital human was given a Scottish accent to resonate more effectively with participants from that region. Based on previous research, such culturally sensitive design decisions can significantly enhance participants' comfort and engagement, potentially leading to more effective therapeutic outcomes [57,58]. Moreover, the input modality is another vital consideration. While speech is often the natural and preferred input method of digital humans, some studies, particularly those conducted in noisy environments like hospitals [36], have opted for touch as the primary mode of interaction. This choice underscores the importance of usability in designing digital humans, emphasizing that how users interact with these entities must be carefully tailored to the specific context in which they are deployed.

Leveraging the advantages of digital humans, we believe that their potential can be extended to a wider range of roles to assist in depression assessment and intervention, especially in helping clients reduce social stigma. For instance, in face-to-face therapy, some clients may bring a support person to the session to help alleviate their anxiety. A digital human can replicate this support function by providing a nonjudgmental and consistent presence, helping clients feel more comfortable and less anxious during their sessions. Another promising role for digital humans is that of a peer specialist. Peer support interventions have been shown to be effective in treating depression, as evidenced by numerous studies [59]. Peer specialists share their own experiences to help clients feel understood and to encourage them to disclose more during the intervention. A digital peer specialist can fulfill this role by simulating these supportive interactions, providing empathetic responses, and sharing relatable experiences that resonate with clients.

RQ2: Design Considerations for Digital Human Implementation

Digital human design choices reflect careful consideration of technological constraints, therapeutic goals, and user contexts. The analysis reveals systematic patterns in how researchers approach fundamental design decisions, from visual appearance to underlying intelligence architectures, with each choice serving specific therapeutic and technical purposes.

Body visibility decisions appear strategically aligned with platform capabilities and therapeutic objectives. Full-body representations dominate VR applications where spatial presence is paramount, while upper-body presentations are preferred for mobile devices where screen space is limited. Head-only implementations are chosen specifically when facial expression fidelity is critical for conveying emotions or avoiding uncanny valley effects [27,32]. This distribution suggests that successful digital human design requires careful consideration of platform affordances and their alignment with therapeutic goals. The predominance of realistic over cartoon-like styling (15 of 17 designs) suggests a preference for human-like appearance that supports therapeutic credibility and user engagement. However, some researchers deliberately chose cartoon styling to avoid uncanny valley effects while maintaining user comfort, indicating that appearance realism must be balanced against potential negative user reactions. Cultural sensitivity emerges as a critical design consideration, with researchers adapting appearance characteristics, accents, and interaction styles to match local population demographics and cultural expectations. These adaptations demonstrate awareness that effective therapeutic relationships require cultural alignment and familiarity.

Interaction modality design reveals sophisticated approaches to creating natural, accessible interfaces. The overwhelming preference for speech-based input (88.2% of implementations) reflects both user expectations for natural conversation and the technological maturity of speech recognition systems. Hence, the conversation ability is one of the main factors that affect closeness perceptions on digital humans [60]. The selective use of touch input in specific contexts, particularly noisy hospital environments, demonstrates adaptive design thinking that prioritizes usability over technological sophistication. This flexibility suggests that effective digital human systems must accommodate diverse deployment contexts and user needs.

Output modality design shows varying levels of sophistication in multimodal communication. While all systems provide speech synthesis and basic behavioral animations, the implementation of empathy narration in only 5 systems suggests that this remains a challenging but valuable design goal. The expanded behavioral repertoires available in 3D environments, including walking, crying, and showing anger, highlight how platform capabilities directly influence design possibilities and therapeutic potential. These variations indicate that output sophistication should align with both technical capabilities and therapeutic requirements.

The distribution of back-end intelligence approaches reveals important insights about current technological capabilities and design priorities. Scripted systems (10/17, 58.8%) dominate in contexts requiring predictable, standardized interactions, particularly for assessment protocols and structured educational content. Rule-based systems (4/17, 23.5%) appear primarily in counseling contexts where greater conversational flexibility is needed while maintaining therapeutic boundaries. This distribution suggests that different therapeutic contexts require different levels of conversational sophistication and autonomy. The continued use of Wizard of Oz techniques (3/17, 17.6%) in recent studies indicates that fully autonomous digital humans may not yet be ready for complex therapeutic interactions, particularly those requiring real-time clinical judgment. This approach allows researchers to explore the therapeutic potential of digital humans while ensuring participant safety and intervention quality. The innovative use of recorded playback in CFT [35] demonstrates how back-end intelligence can be tailored to specific therapeutic mechanisms, using participants' own recorded behaviors to create personalized therapeutic content.

The design of digital humans for depression support requires balancing multiple factors: Visual appearance needs to be culturally adapted to the target population and matched to platform capabilities, and within these constraints, should be realistic enough for credibility while avoiding uncanny valley effects. Interaction design should prioritize natural speech input where feasible but adapt to environmental constraints. Back-end intelligence approaches must be selected based on therapeutic context, with scripted systems providing safety and standardization for assessments, while rule-based and Wizard of Oz techniques enable more flexible therapeutic interactions. Ultimately, successful digital human implementation depends on aligning design choices with specific therapeutic goals, technological capabilities, and user needs rather than pursuing technological sophistication for its own sake.

Digital Human in Building Therapeutic Alliance

The therapeutic alliance, characterized by a collaborative and trusting relationship between a health provider and a client, is a cornerstone of effective psychotherapy [61]. Traditionally, this alliance is built through face-to-face interactions, where empathy, understanding, and mutual respect foster a sense of safety and connection. With the advent of digital humans in therapeutic settings, there is a growing interest in understanding how these virtual entities can contribute to building and maintaining a therapeutic alliance [62].

In this scoping review, multiple studies [23,36,37,39] have demonstrated the potential of digital humans to establish trustworthy and comfortable relationships with clients. These findings underscore the viability of digital humans as supportive agents in mental health care. Several key factors contribute to the successful establishment of therapeutic alliances with digital humans. First is empathy and emotional responsiveness. The ability to empathize is crucial in building a therapeutic alliance [63]. Digital humans, as embodied virtual agents, can combine facial expressions, body movements, and emotional narration to convey emotions accurately. In the study by Ring et al [37], participants felt that their feelings were understood because the digital human provided appropriate feedback. Similar results were observed in the experiment by Philip et al [39], where participants also reported a sense of being understood. These findings highlight the importance of empathetic interactions in fostering a strong therapeutic alliance. Another critical factor is the nonjudgmental presence of digital humans. This presence is created through the careful design and control of the content expressed by the digital human. The interactive content of digital humans applied in all of the 17 included studies was designed or evaluated by experts before being applied in experiments. Although this approach limits the range of user interactions, it ensures the safety and appropriateness of the interactions. By providing a nonjudgmental and safe space, digital humans can help clients feel more comfortable and open, thereby strengthening the therapeutic alliance. In addition, compared to face-to-face interactions, digital humans offer customizable support tailored to the individual's personality and needs. This personalized attention can make people with depression feel valued and understood, increasing their motivation to engage in therapy and fostering a deeper sense of trust.

Digital humans have demonstrated significant potential in building therapeutic alliances with individuals experiencing depression. By providing empathetic, consistent, and personalized support, digital humans can help overcome barriers related to stigma, availability, and accessibility.

Ethical Considerations and Potential Risks

While this scoping review demonstrates the promising potential of digital humans in depression management, implementation raises critical ethical concerns requiring careful attention. Privacy and data security present immediate risks, as digital human systems collect highly sensitive information, including verbal responses [35], behavioral patterns [39], and potentially video or audio recordings during vulnerable therapeutic moments [37]. Many reviewed studies provided limited detail about data management practices, yet these systems often rely on cloud-based infrastructure and third-party services that create multiple exposure points for sensitive mental health data. Biometric information, such as voice recordings and facial expressions, is inherently difficult to anonymize and could potentially reidentify individuals even after traditional identifiers are removed. Future implementations should ensure robust encryption, secure storage protocols compliant with relevant regulations, transparent disclosure of data practices, and clear policies regarding third-party access and data retention.

The accessibility advantages of digital humans paradoxically create risks of overreliance and inappropriate use. Recent news has shown that individuals may incorrectly view these tools as complete replacements for human therapists rather than complementary supports [64], potentially delaying access to necessary human intervention for severe depression, suicidal ideation, or complex presentations. Current autonomous systems have limited capacity to recognize and appropriately respond to acute crises, and users may form parasocial relationships with digital humans without recognizing fundamental limitations compared with human therapeutic relationships. Clear communication about system boundaries is essential, alongside robust escalation mechanisms to human providers when needed. Additionally, questions of accountability remain unresolved—when adverse outcomes occur, responsibility distributed among developers, health care providers, and users requires clear frameworks that protect vulnerable populations while enabling innovation. Equity and access concerns threaten to undermine the democratizing potential of digital humans. Implementation requires technological access (appropriate devices, reliable internet, and digital literacy) that populations at risk for depression—including low-income individuals, older adult persons, and those in rural areas—may lack. Only 25% (5/20) [23,25,33,34,38] of reviewed studies occurred in everyday environments, suggesting limited real-world accessibility evidence. Without deliberate policy intervention ensuring equitable access through public health systems and subsidized programs, these technologies risk becoming available primarily through private markets, widening rather than closing mental health care gaps.

Moving forward, ethical implementation requires privacy-by-design approaches with comprehensive data protection, transparent communication of system limitations and appropriate use cases, inclusive development ensuring diverse representation in training data and validation studies, and hybrid care models positioning digital humans as adjuncts rather than replacements for human providers. Regulatory frameworks must establish appropriate validation standards and postmarket surveillance requirements, while sustainable funding models must prioritize access for underserved populations. As this technology advances beyond proof-of-concept demonstrations, the field should evaluate not only what digital humans can do, but what they should do, ensuring that the pursuit of innovation maintains focus on user well-being, autonomy, safety, and equitable access across all populations who might benefit from mental health support.

Suggestions for Future Research

A total of 3 promising research opportunities emerged from this review that offer significant potential to advance the field of digital humans in depression management.

First, the current technological landscape reveals substantial opportunities for enhancement that could transform therapeutic capabilities. Future research would benefit from exploring the integration of physiological monitoring (heart rate variability, galvanic skin response, and eye tracking) with digital human systems to enable real-time emotional state detection and adaptive responses. Given that current systems rely predominantly on scripted or rule-based approaches, there is potential for incorporating enhanced natural language processing and emotional artificial intelligence capabilities while maintaining therapeutic safety standards. Cross-platform optimization research presents exciting opportunities, as current use of digital humans has appeared across mobile platforms (smartphones and tablets) [25,38], stationary platforms (large display systems and desktop monitors) [21,27], and VR headsets [35,40], suggesting the need for thoughtful platform-specific design guidelines and seamless integration across devices to maintain therapeutic continuity.

Building on the demonstrated success of digital humans in traditional clinical roles, research could explore innovative therapeutic applications that leverage their unique capabilities. The encouraging success of digital humans as "actors" representing emotions and thoughts suggests promising potential for novel approaches, including digital peer specialists who can share lived experiences, family therapy facilitators who manage complex group dynamics, and group intervention leaders who ensure equitable participation. Prevention and early intervention applications present valuable research opportunities, as current studies focus primarily on individuals already experiencing depression rather than exploring applications with at-risk populations, such as adolescents, caregivers, or individuals with chronic medical conditions. Research could also investigate how digital humans can serve as effective mediators between counselors and clients, facilitating communication and building trust through their demonstrated ability to reduce stigma and encourage self-disclosure.

Finally, research would greatly benefit from investigating how digital humans can be effectively integrated into existing health care workflows, including exploring optimal combinations of digital and human therapist interactions, developing smooth hand-off protocols between digital and human providers, and establishing data-sharing mechanisms that maintain continuity of care. Cost-effectiveness analyses offer valuable opportunities to demonstrate the economic benefits of digital human interventions compared with traditional care models, while implementation research could systematically explore facilitators and address barriers to adoption across diverse clinical settings, from large hospital systems to community health centers.

Besides these research opportunities, we also recognize the need for concrete methodological guidance to advance the field beyond its current proof-of-concept stage. Future studies should use adequate sample sizes appropriate to their RQs: proof-of-concept studies should include a minimum of 30 - 50 participants for preliminary feasibility assessment, while comparative effectiveness trials require 100 - 150 per arm to detect medium effect sizes with adequate power. Researchers should adopt standardized outcome measures to enable meta-analyses and cross-study comparisons, including PHQ-9 or Beck Depression Inventory-II for depression symptom severity, the Working Alliance Inventory adapted for digital humans to assess therapeutic alliance, the System Usability Scale for usability evaluation, and systematic adverse event monitoring. Additionally, we suggest that future research include more detailed documentation of the digital human design process to enhance transparency and replicability. As noted in our limitations, most reviewed studies provided only final design descriptions without elaborating on the underlying rationale, iterative decisions, or user-centered design methodologies used.

Practical Implications

User-Centered Design Principles

The review highlights that effective digital human systems are grounded in collaborative design processes between clinicians and developers. To ensure therapeutic relevance and usability:

- Clinicians should be involved early to define therapeutic goals, patient needs, and appropriate boundaries.
- Developers should guide feasibility discussions and apply user experience principles tailored to mental health contexts.
- Ongoing dialogue should promote mutual understanding—clinicians educate on therapeutic aspects of the project, while developers clarify what current technologies can and cannot achieve.

This interdisciplinary co-design ensures that systems are both clinically sound and technically viable.

Ethical Implementation Guidelines

The deployment of digital humans in mental health care raises important ethical considerations that require careful attention during study. Based on the review findings, we suggest:

- Study protocols should prioritize transparency about the interactions between participants and digital humans, especially the data collection practices.
- Participants should be clearly informed about the capabilities and limitations of digital human systems, including their role as therapeutic tools rather than replacements for human care.
- Implementation frameworks should establish clear boundaries and expectations from the outset—clinicians define appropriate therapeutic limits and user relationship parameters, while developers implement technical safeguards that support these ethical boundaries.

Integration Into Clinical Workflows

Successful implementation of digital humans in depression care requires strategic planning across multiple dimensions:

- Design the digital human's appearance and language to reflect target user population characteristics (eg, facial features, accent, and ethnic background).
- Assess the physical environment where deployment will occur and select interaction modalities appropriate to the setting (eg, touch-based input for noisy environments and speech input for quiet, private spaces).
- Start with well-defined tasks where scripted systems have proven effective: standardized questionnaire administration (PHQ-9), protocol-based interviews, and psychoeducational content delivery.
- Provide training for clinical staff on system capabilities, limitations, and appropriate use cases.
- Implement feedback mechanisms to continuously improve system performance based on clinician and patient input.

Limitations

This scoping review has some limitations. To make our review feasible, we used a relatively narrow literature search approach, which may have introduced selection bias. The restriction to studies published in English could have excluded relevant research, and due to the various terms used to describe digital humans in the literature, our search terms might not have covered all relevant papers.

The studies included in this review exhibited considerable heterogeneity in terms of design, methodology, and intervention protocols. This makes it challenging to draw definitive conclusions or directly compare outcomes across different studies. Critically, the evidence base is dominated by proof-of-concept studies and small-sample trials. Specifically, 55% of studies (n=11) [23-26,28-30,33-35,37] included fewer than 50 participants, with only 20% (n=4) [31,36,38,39] enrolling over 100 participants. This prevalence of small-scale exploratory research limits the statistical power and generalizability of findings. While these developmental studies provide valuable preliminary evidence, they are insufficient for making definitive claims about therapeutic efficacy. The differences in study designs, sample sizes, and measurement tools underscore that digital humans for depression management remain an emerging field requiring substantial methodological advancement before strong clinical recommendations can be made.

Another limitation identified in the included studies is the lack of detailed information on the design processes for digital humans in the reviewed studies. Most studies provided descriptions of the functions and appearances of digital humans but did not elaborate on the underlying design choices or rationales. This lack of transparency makes it difficult to understand how specific design decisions might have influenced study outcomes and limits our ability to evaluate the replicability and effectiveness of different digital human implementations. As with most systematic literature reviews, potential publication bias is also a concern, as unpublished studies or those with negative results might not have been included in this review.

Conclusions

This scoping review of 20 studies systematically addressed how digital humans are used in depression management and their design considerations. Regarding usage (RQ1), digital humans demonstrate versatility across assessment (9/20, 45%) through questionnaire administration, interviews, and interactive tasks, and intervention (13/20, 65%) through CBT, psychoeducation, and innovative therapies, assuming roles as interviewers, facilitators, counselors, educators, and actors. For design considerations (RQ2), successful implementations strategically align appearance with platform capabilities (full-body for VR and upper-body for mobile), prioritize realistic styling and cultural sensitivity, use speech-based input (15/17, 88.2%), and use back-end intelligence ranging from scripted (10/17, 58.8%) to rule-based systems. The use of digital humans now stands at a juncture where it not only enables bidirectional conversations but also significantly enriches interactions within the domain of depression management. This advancement signals a paradigm shift toward fostering deeper connections, tailored assistance, and broadening the horizons of accessibility, surpassing what traditional therapeutic frameworks could offer. Our scoping review has charted the pioneering deployment of digital humans across the dual spectrums of assessment and intervention, casting a spotlight on their transformative potential to amplify the reach and resonance of mental health interventions.

Funding

The project leading to this review has received funding from Ministry of Business, Innovation and Employment (PROP-84294-CATAI-UOA).

Conflicts of Interest

None declared.

Checklist 1
PRISMA-ScR checklist.
[PDF File, 520 KB - mental_v13i1e79954_app1.pdf]

References

1. Depressive disorder (depression). World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/depression/?gad_source=1&gclid=CjwKCAiAxaCvBhBaEiwAvsLmWLmQ_ozx9P-8vh1GORQUMfs85ezcBENdx8Az5e6SMhZGBVP64sGGlBoCqkwQAvD_BwE [accessed 2025-12-22]
2. Saarni SI, Suvisaari J, Sintonen H, et al. Impact of psychiatric disorders on health-related quality of life: general population survey. Br J Psychiatry 2007 Apr;190:326-332. [doi: 10.1192/bjp.bp.106.025106] [Medline: 17401039]
3. Cuijpers P, Karyotaki E, Eckshtain D, et al. Psychotherapy for depression across different age groups: a systematic review and meta-analysis. JAMA Psychiatry 2020 Jul 1;77(7):694-702. [doi: 10.1001/jamapsychiatry.2020.0164] [Medline: 32186668]
4. Zhou X, Hetrick SE, Cuijpers P, et al. Comparative efficacy and acceptability of psychotherapies for depression in children and adolescents: a systematic review and network meta-analysis. World Psychiatry 2015 Jun;14(2):207-222. [doi: 10.1002/wps.20217] [Medline: 26043339]
5. Campbell LF, Norcross JC, Vasquez MJT, Kaslow NJ. Recognition of psychotherapy effectiveness: the APA resolution. Psychotherapy (Chic) 2013 Mar;50(1):98-101. [doi: 10.1037/a0031817] [Medline: 23505985]

6. McHugh RK, Whitton SW, Peckham AD, Welge JA, Otto MW. Patient preference for psychological vs pharmacologic treatment of psychiatric disorders: a meta-analytic review. J Clin Psychiatry 2013 Jun;74(6):595-602. [doi: 10.4088/JCP.12r07757] [Medline: 23842011]
7. Karyotaki E, Smit Y, Holdt Henningsen K, et al. Combining pharmacotherapy and psychotherapy or monotherapy for major depression? A meta-analysis on the long-term effects. J Affect Disord 2016 Apr;194:144-152. [doi: 10.1016/j.jad.2016.01.036] [Medline: 27262637]
8. Hoermann S, McCabe KL, Milne DN, Calvo RA. Application of synchronous text-based dialogue systems in mental health interventions: systematic review. J Med Internet Res 2017 Jul 21;19(8):e267. [doi: 10.2196/jmir.7023] [Medline: 28784594]
9. Burden D, Savin-Baden M. Virtual Humans: Today and Tomorrow: Chapman & Hall/CRC; 2019.
10. Kocaballi AB, Berkovsky S, Quiroz JC, et al. The personalization of conversational agents in health care: systematic review. J Med Internet Res 2019 Nov 7;21(11):e15360. [doi: 10.2196/15360] [Medline: 31697237]
11. Ma T, Sharifi H, Chattopadhyay D. Virtual humans in health-related interventions. 2019 May 2 Presented at: CHI ’19: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems; Glasgow Scotland UK p. 1-6 URL: https://dl.acm.org/doi/10.1145/3290607.3312853 [accessed 2025-12-22] [doi: 10.1145/3290607.3312853]
12. Chattopadhyay D, Ma T, Sharifi H, Martyn-Nemeth P. Computer-controlled virtual humans in patient-facing systems: systematic review and meta-analysis. J Med Internet Res 2020 Jul 30;22(7):e18839. [doi: 10.2196/18839] [Medline: 32729837]
13. Li J, Theng YL, Foo S. Game-based digital interventions for depression therapy: a systematic review and meta-analysis. Cyberpsychol Behav Soc Netw 2014 Aug;17(8):519-527. [doi: 10.1089/cyber.2013.0481] [Medline: 24810933]
14. Moshe I, Terhorst Y, Philippi P, et al. Digital interventions for the treatment of depression: a meta-analytic review. Psychol Bull 2021 Aug;147(8):749-786. [doi: 10.1037/bul0000334] [Medline: 34898233]
15. Wang F, Preininger A. AI in health: state of the art, challenges, and future directions. Yearb Med Inform 2019 Aug;28(1):16-26. [doi: 10.1055/s-0039-1677908] [Medline: 31419814]
16. Arksey H, O’Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
17. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143. [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
18. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
19. Traum D. Models of culture for virtual human conversation. 2009 Presented at: Universal Access in Human-Computer Interaction Applications and Services: 5th International Conference, UAHCI 2009, held as part of HCI International 2009; Jul 19-24, 2009; San Diego, CA p. 434-440 URL: https://link.springer.com/chapter/10.1007/978-3-642-02713-0_46 [accessed 2025-12-22] [doi: 10.1007/978-3-642-02713-0_46]
20. Cuijpers P. Indirect prevention and treatment of depression: an emerging paradigm? Clin Psychol Eur 2021 Dec;3(4):e6847. [doi: 10.32872/cpe.6847] [Medline: 36398290]
21. Jaiswal S, Valstar M, Kusumam K, Greenhalgh C. Virtual human questionnaire for analysis of depression, anxiety and personality. 2019 Jul Presented at: IVA ’19: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents; Paris France p. 81-87. [doi: 10.1145/3308532.3329469]
22. Egede JO, Price D, Krishnan DB, et al. Design and evaluation of virtual human mediated tasks for assessment of depression and anxiety. 2021 Sep 14 Presented at: IVA ’21: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents; Virtual Event Japan p. 52-59. [doi: 10.1145/3472306.3478361]
23. Wolters MK, Tâtar AS, Matu S, et al. eHealth support for people with depression in the community: a case study series. Presented at: HCIK ’15: Proceedings of HCI Korea; Dec 10-12, 2014 p. 138-144 URL: https://dl.acm.org/doi/10.5555/2729485.2729507 [accessed 2025-12-22]
24. Baghaei N, Stemmet L, Khaliq I, et al. Designing individualised virtual reality applications for supporting depression: a feasibility study. 2021 Jun 8 Presented at: EICS ’21; Virtual Event Netherlands p. 6-11. [doi: 10.1145/3459926.3464761]
25. Luerssen MH, Hawke T. Virtual agents as a service: applications in healthcare. 2018 Presented at: Proceedings of the 18th International Conference on Intelligent Virtual Agents; Nov 5-8, 2018 p. 107-112. [doi: 10.1145/3267851.3267858]
26. Takemoto A, Aispuriete I, Niedra L, Dreimane LF. Depression detection using virtual avatar communication and eye tracking. J Eye Mov Res 2023;16(2). [doi: 10.16910/jemr.16.2.6] [Medline: 38075672]
27. Bresó A, Martínez‐Miranda J, Botella C, Baños RM, García‐Gómez JM. Usability and acceptability assessment of an empathic virtual agent to prevent major depression. Expert Systems 2016 Aug;33(4):297-312 [FREE Full text] [doi: 10.1111/exsy.12151]
28. Takemoto A, Aispuriete I, Niedra L, Dreimane LF. Differentiating depression using facial expressions in a virtual avatar communication system. Front Digit Health 2023;5:1080023. [doi: 10.3389/fdgth.2023.1080023] [Medline: 36969955]
29. DeVault D, Georgila K, Artstein R, et al. Verbal indicators of psychological distress in interactive dialogue with a virtual human. : Association for Computational Linguistics; 2013 Presented at: Proceedings of the SIGDIAL 2013 Conference; Aug 22-24, 2013 p. 193-202 URL: https://aclanthology.org/W13-4032/ [accessed 2025-12-22]

30. Ashrafi N, Neuhaus V, Vona F, Peperkorn NL, Shiban Y, Voigt-Antons JN. Effect of external characteristics of a virtual human being during the use of a computer-assisted therapy tool. Presented at: International Conference on Human-Computer Interaction; Jun 29 to Jul 4, 2024 p. 3-21 URL: https://link.springer.com/chapter/10.1007/978-3-031-60428-7_1 [accessed 2025-12-22] [doi: 10.1007/978-3-031-60428-7_1]
31. Wu EHK, Gao TY, Chung CR, Chen CC, Tsai CF, Yeh SC. Mobile virtual assistant for multi-modal depression-level stratification. IEEE Trans Affective Comput 2024;16(2):611-623. [doi: 10.1109/TAFFC.2024.3451114]
32. Kocur M, Dechant M, Wolff C, et al. Computer-assisted avatar-based treatment for dysfunctional beliefs in depressive inpatients: a pilot study. Front Psychiatry 2021;12:608997. [doi: 10.3389/fpsyt.2021.608997] [Medline: 34335319]
33. Burton C, Szentagotai Tatar A, McKinstry B, et al. Pilot randomised controlled trial of Help4Mood, an embodied virtual agent-based system to support treatment of depression. J Telemed Telecare 2016 Sep;22(6):348-355. [doi: 10.1177/1357633X15609793] [Medline: 26453910]
34. Shamekhi A, Bickmore T, Lestoquoy A, Negash L, Gardiner P. Blissful agents: adjuncts to group medical visits for chronic pain and depression. Presented at: Intelligent Virtual Agents: 16th International Conference, IVA 2016; Sep 20-23, 2016; Los Angeles, CA p. 433-437. [doi: 10.1007/978-3-319-47665-0_49]
35. Halim I, Stemmet L, Hach S, et al. Individualized virtual reality for increasing self-compassion: evaluation study. JMIR Ment Health 2023 Oct 2;10:e47617. [doi: 10.2196/47617] [Medline: 37782537]
36. Bickmore TW, Mitchell SE, Jack BW, Paasche-Orlow MK, Pfeifer LM, Odonnell J. Response to a relational agent by hospital patients with depressive symptoms. Interact Comput 2010 Jul 1;22(4):289-298. [doi: 10.1016/j.intcom.2009.12.001] [Medline: 20628581]
37. Ring L, Bickmore T, Pedrelli P. An affectively aware virtual therapist for depression counseling. Semantic Scholar. URL: https://www.semanticscholar.org/paper/An-Affectively-Aware-Virtual-Therapist-for-Ring-Bickmore/004ec53d1f12cc4c0a7c809bf3b7acaee2180fd9 [accessed 2025-12-22]
38. Shamekhi A, Bickmore T, Lestoquoy A, Gardiner P. Augmenting group medical visits with conversational agents for stress management behavior change. 2017 Presented at: PERSUASIVE Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors: 12th International Conference, PERSUASIVE 2017; Apr 4-6, 2017 URL: https://www.researchgate.net/publication/314712342_Augmenting_Group_Medical_Visits_with_Conversational_Agents_for_Stress_Management_Behavior_Change [accessed 2025-12-30]
39. Philip P, Micoulaud-Franchi JA, Sagaspe P, et al. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. Sci Rep 2017 Feb 16;7:42656. [doi: 10.1038/srep42656] [Medline: 28205601]
40. Hidding M, Veling W, Pijnenborg GHM, van der Stouwe ECD. A single-session VR intervention addressing self-compassion and self-criticism with and without perspective change: results of a randomized controlled experiment. Behav Res Ther 2024 Feb;173:104466. [doi: 10.1016/j.brat.2023.104466] [Medline: 38141543]
41. Hofmann SG, Asnaani A, Vonk IJJ, Sawyer AT, Fang A. The efficacy of cognitive behavioral therapy: a review of meta-analyses. Cognit Ther Res 2012 Oct 1;36(5):427-440. [doi: 10.1007/s10608-012-9476-1] [Medline: 23459093]
42. Gilbert P. Introducing compassion-focused therapy. Adv Psychiatr Treat 2009 May;15(3):199-208. [doi: 10.1192/apt.bp.107.005264]
43. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
44. Moullec G, Maïano C, Morin AJS, Monthuy-Blanc J, Rosello L, Ninot G. A very short visual analog form of the Center for Epidemiologic Studies Depression Scale (CES-D) for the idiographic measurement of depression. J Affect Disord 2011 Feb;128(3):220-234. [doi: 10.1016/j.jad.2010.06.006] [Medline: 20609480]
45. Paykel ES. The clinical interview for depression. Development, reliability and validity. J Affect Disord 1985 Jul;9(1):85-96. [doi: 10.1016/0165-0327(85)90014-x] [Medline: 3160752]
46. Depressive Disorders: DSM-5 Selections: American Psychiatric Association; 2015.
47. Li YJ, Huang A, Sanku BS, He JS. Designing an empathy training for depression prevention using virtual reality and a preliminary study. 2023 Presented at: 2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW); Mar 25-29, 2023; Shanghai, China p. 44-52. [doi: 10.1109/VRW58643.2023.00015]
48. Butler AC, Chapman JE, Forman EM, Beck AT. The empirical status of cognitive-behavioral therapy: a review of meta-analyses. Clin Psychol Rev 2006 Jan;26(1):17-31. [doi: 10.1016/j.cpr.2005.07.003] [Medline: 16199119]
49. Clark DA. Cognitive restructuring. In: The Wiley Handbook of Cognitive Behavioral Therapy: Wiley Blackwell; 2013:1-22. [doi: 10.1002/9781118528563]
50. Cuijpers P, van Straten A, Warmerdam L. Behavioral activation treatments of depression: a meta-analysis. Clin Psychol Rev 2007 Apr;27(3):318-326. [doi: 10.1016/j.cpr.2006.11.001] [Medline: 17184887]
51. Waller H, Garety PA, Jolley S, et al. Low intensity cognitive behavioural therapy for psychosis: a pilot study. J Behav Ther Exp Psychiatry 2013 Mar;44(1):98-104. [doi: 10.1016/j.jbtep.2012.07.013] [Medline: 22940787]
52. Tursi MDS, Baes CVW, Camacho FDB, Tofoli SDC, Juruena MF. Effectiveness of psychoeducation for depression: a systematic review. Aust N Z J Psychiatry 2013 Nov;47(11):1019-1031. [doi: 10.1177/0004867413491154]

53. Stratou G, Scherer S, Gratch J, Morency LP. Automatic nonverbal behavior indicators of depression and PTSD: the effect of gender. J Multimodal User Interfaces 2015 Mar;9(1):17-29. [doi: 10.1007/s12193-014-0161-4]
54. Dahlbäck N, Jönsson A, Ahrenberg L. Wizard of Oz studies: why and how. Presented at: IUI ’93: Proceedings of the 1st International Conference on Intelligent User Interfaces; Jan 4-7, 1993; Orlando, FL p. 193-200. [doi: 10.1145/169891.169968]
55. Lucas GM, Gratch J, King A, Morency LP. It’s only a computer: virtual humans increase willingness to disclose. Comput Human Behav 2014 Aug;37:94-100. [doi: 10.1016/j.chb.2014.04.043]
56. Loveys K, Sagar M, Broadbent E. The effect of multimodal emotional expression on responses to a digital human during a self-disclosure conversation: a computational analysis of user language. J Med Syst 2020 Jul 22;44(9):020-01624. [doi: 10.1007/s10916-020-01624-4] [Medline: 32700060]
57. Sun H. Cross-Cultural Technology Design: Creating Culture-Sensitive Technology for Local Users: Oxford University Press; 2012. [doi: 10.1093/ACPROF:OSO/9780199744763.001.0001]
58. Mohamed H, Al-Lenjawi B, Amuna P, Zotor F, Elmahdi H. Culturally sensitive patient-centred educational programme for self-management of type 2 diabetes: a randomized controlled trial. Prim Care Diabetes 2013 Oct;7(3):199-206. [doi: 10.1016/j.pcd.2013.05.002] [Medline: 23830727]
59. Pfeiffer PN, Heisler M, Piette JD, Rogers MAM, Valenstein M. Efficacy of peer support interventions for depression: a meta-analysis. Gen Hosp Psychiatry 2011;33(1):29-36. [doi: 10.1016/j.genhosppsych.2010.10.002] [Medline: 21353125]
60. Loveys K, Hiko C, Sagar M, Zhang X, Broadbent E. “I felt her company”: a qualitative study on factors affecting closeness and emotional support seeking with an embodied conversational agent. Int J Hum Comput Stud 2022 Apr;160:102771. [doi: 10.1016/j.ijhcs.2021.102771]
61. Martin DJ, Garske JP, Davis MK. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. J Consult Clin Psychol 2000 Jun;68(3):438-450. [doi: 10.1037/0022-006X.68.3.438] [Medline: 10883561]
62. D’Alfonso S, Lederman R, Bucci S, Berry K. The digital therapeutic alliance and human-computer interaction. JMIR Ment Health 2020 Dec 29;7(12):e21895. [doi: 10.2196/21895] [Medline: 33372897]
63. Feller CP, Cottone RR. The importance of empathy in the therapeutic alliance. The Journal of Humanistic Counseling, Education and Development 2003 Mar;42(1):53-61. [doi: 10.1002/j.2164-490X.2003.tb00168.x]
64. “Sliding into an abyss”: experts warn over rising use of AI for mental health support. The Guardian. 2025 Aug 30. URL: https://www.theguardian.com/society/2025/aug/30/therapists-warn-ai-chatbots-mental-health-support [accessed 2025-12-22]

Abbreviations

CAT-DB: computer-assisted avatar-based treatment for dysfunctional beliefs
CBT: cognitive behavioral therapy
CES-D-VAS-VS: Center for Epidemiologic Studies Depression Scale - Visual Analog Scale - Very Short version
CFT: compassion-focused therapy
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
PHQ-9: Patient Health Questionnaire-9
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RQ: research question
TAU: treatment as usual
VR: virtual reality

Edited by A Lustig; submitted 02.Jul.2025; peer-reviewed by M Adnan, O Ajileye; revised version received 30.Sep.2025; accepted 06.Oct.2025; published 05.Jan.2026.

Please cite as:
Cao J, Gao W, Wen R, Li C, Hoermann S, Baghaei N, Billinghurst M
Digital Humans for Depression Assessment and Intervention Support: Scoping Review
JMIR Ment Health 2026;13:e79954
URL: https://mental.jmir.org/2026/1/e79954
doi:10.2196/79954

Mood Monitoring, Mood Tracking, and Ambulatory Assessment Interventions in Depression and Bipolar Disorder: Systematic Review and Meta-Analysis of Randomized Controlled Trials

Laurence Astill Wright[1,2], MRCPsych; Georgina Shajan[1]; Daljit Purewal[1]; Jonathan Stone[3], BMBS, BSc (Hons); Madiha Majid[1,4], MRCPsych; Boliang Guo[1,5,6,7], PhD; Richard Morriss[1,5,6,7], PhD

[1]Institute of Mental Health, University of Nottingham, Jubilee Campus, Triumph Road, Nottingham, United Kingdom
[2]Centre for Academic Mental Health, Population Health Sciences, University of Bristol, Bristol, United Kingdom
[3]Avon and Wiltshire Mental Health Partnership NHS Trust, Bristol, United Kingdom
[4]Coventry and Warwickshire Partnership NHS Trust, Coventry, United Kingdom
[5]NIHR ARC East Midlands, University of Nottingham, Nottingham, United Kingdom
[6]Nottingham NIHR Biomedical Research Centre, University of Nottingham, Nottingham, United Kingdom
[7]NIHR MindTech Medical Technology Collaborative, University of Nottingham, Nottingham, United Kingdom

Corresponding Author:
Laurence Astill Wright, MRCPsych
Institute of Mental Health, University of Nottingham, Jubilee Campus, Triumph Road, Nottingham, United Kingdom

Abstract

Background: Mood monitoring is widely used by people with depression and bipolar disorder (BD) to prevent relapse and improve insight into their condition, but it is unclear if these interventions have an impact on symptoms and for whom. As the capacity for passive mood monitoring increases, it is vital to improve our understanding of frequent mood assessment.

Objective: This systematic review and meta-analysis assessed the effect of mood monitoring interventions in people with depression and BD to decrease relapse risk and symptoms of depression and mania.

Methods: We conducted a systematic review and meta-analysis (PROSPERO, International Prospective Register of Systematic Reviews: CRD42023396473) and reported results according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) guidelines. Randomized controlled trials with clinically important follow-up periods were identified via multiple database searches and rated for risk of bias using the Cochrane Risk of Bias tool. The primary outcomes were symptoms of depression and mania. Available data were pooled to calculate standardized mean differences (SMDs) for the primary outcomes: severity of depression, bipolar depression, and mania/hypomania.

Results: We included 8 trials of 1230 participants and 6 different mood monitoring protocols. In BD, meta-analysis found a small but not statistically significant effect of mood monitoring interventions on decreasing mania symptoms (6 comparisons, n=873; SMD 0.16, 95% CI−0.34 to 0.01; *P*=.06) and no effect on bipolar depression (6 comparisons, n=873; SMD −0.08, 95% CI −0.31 to 0.15; *P*=.02). In depression, we found a small effect in decreasing symptoms of depression of borderline statistical significance at 12 months (2 comparisons, n=262; SMD −0.25, 95% CI −0.49 to 0.00; *P*=.05) but not at 6 months (2 comparisons, n=268; SMD −0.21, 95% CI −0.54 to 0.12; *P*=.21). There was an absence of evidence on the effect of mood monitoring on decreased relapse rates or readmission rates. Studies had a low risk of bias. There was no evidence on mood monitoring through ecological momentary assessment.

Conclusions: Overall mood monitoring interventions do not increase or decrease mood symptoms in people with BD, nor is there robust evidence of such effects in people with unipolar depression. Further research is merited on different forms of mood monitoring and to determine under what circumstances mood monitoring might have beneficial or adverse effects. These results initially suggest that ambulatory assessment does not induce large placebo effects or significantly negatively or positively affect mood, and thus that mood monitoring may be an appropriate outcome measure for research or for clinical practice.

(JMIR Ment Health 2026;13:e84020) doi:10.2196/84020

KEYWORDS

bipolar; depression; ecological momentary assessment; mood-tracking; mood-monitoring; self-monitoring

Introduction

Many people with bipolar disorder (BD) and depression track their mood symptoms over time, and there are multiple tools available freely on the web to do this [1-4]. A recent survey of people with BD found that 41.6% of participants reported using a self-management app related to mood or sleep [5,6]. Mood monitoring is widely used, specifically by people with mood disorders [7], for example, Bipolar UK's mood tracking app has greater than 10,000 downloads on Android alone [8]. Mood tracking is one of the most common features of mental health smartphone apps—previous reviews have noted that mood/behavior tracking is present in over half of these apps [9-11]. Many smartphones include in-built self-tracking functions for health [12], and many studies are incorporating mood monitoring as a method of clinical outcome [13] and testing them as interventions [14].

Traditionally, mood monitoring is done place using paper-based charting [15]. However, many people with BD and depression prefer digital methods as they are convenient and store easily accessible records of mood, allowing people to more easily look back and identify patterns of improvement or worsening of symptoms [8]. Digital methods, where individuals record their mood in the moment, may also decrease recall bias, so there might be greater accuracy in charting and plotting mood rather than retrospective completion of data every few weeks [7]. The technology used in mood tracking and ambulatory assessment is wide-ranging, and some of the descriptive terms used throughout this paper are nuanced, often with overlapping definitions. Because of this, we have listed these important terms in Table 1 below.

Table . Definitions of forms of mood-monitoring and related terms.

Term	Definition
Self-monitoring	The appraisal and recording of one's current state, can include mood.
Mood tracking/mood monitoring	Regular recording of one's mood over a period of time. This can be done digitally on a device or analog using pen and paper charting.
Ambulatory assessment	Wide group of digital methods recording data on the user in real time and in natural settings. Includes mood tracking/monitoring, remote measurement technology, and ecological momentary assessment.
Active data collection	Users input information about their own current state.
Passive data collection	Behavioral data is automatically recorded via technology.
Remote measurement technology	Wearable devices record passive data.
Ecological momentary assessment	Intensive "in the moment" self-reporting by the user, for example, multiple times per day.

Mood monitoring can be used as an intervention (both in randomized controlled trials [RCTs] and nonrandomized studies) and also as a method of ascertaining outcome (both in RCTs and nonrandomized studies). Passive data collection may reduce the burden of data completion and remind the participant less frequently about their mood [16,17]. Some mood monitoring may combine active and passive monitoring. For example, passive monitoring of certain activity or behavior may trigger active data collection from the user when there is a preset level of change in this activity/behavior. In depression, activity may be reduced, and in mania, it may increase [18]. Other forms of mood monitoring might randomly request the participant to actively complete data on mood without any passive monitoring [17]. These technological advances may provide new utility to a relatively old intervention methodology. However, there is a need to assess whether these newer approaches to mood monitoring have benefits or harms as well, and so we included all of these approaches in this review.

There is evidence that increasing awareness of mood fluctuations can improve insight, and the identification of early warning signs can prevent relapse in depression and BD [19]. This raises the question of whether mood tracking can have any direct clinical effects, either positive or negative [20]. Currently, it is unclear if mood monitoring or mood tracking as an intervention is effective in reducing symptom severity or in preventing relapse. It is also possible that mood monitoring interventions have negative effects on mood [21,22] or lead to a response bias whereby users complete the same score despite mood altering in response to being asked the same questions repeatedly. As the capacity for mood monitoring through digital assessment increases and these methods are used increasingly as assessment methods in research, it is vital to improve our understanding of frequent mood assessment [23].

Some people with BD report that mood monitoring helps them to reduce relapse risk, for example, through greater awareness of their current mental state, while others report it worsens their mood, for example, by reminding them of their mood problems, so they consider mood monitoring a burden [16,22]. Others report that it is relatively simple to carry out in their day-to-day lives [1]. Mood monitoring may represent an intervention that is usable, acceptable, and easy for individuals to implement in their lives, and it can be coupled quite easily with simple psychological interventions such as psychoeducation [24]. Questions remain, however, about definitive efficacy and potential for harm from mood monitoring alone.

The potential for efficacy or adverse events is particularly important as RCTs and observational studies increasingly move towards digital combinations of passive and active monitoring or ambulatory assessment outcomes [17,25]. It is thus integral to know whether the method of assessment itself may carry any therapeutic benefits or any adverse effects. Any such risk or benefit might bias any outcome assessment of other interventions in trials, potentially enhancing or obscuring any true benefit or suggesting a false benefit through measurement methods used in the trial rather than from the intervention itself [26]. Furthermore, if there were adverse effects, ambulatory assessment protocols would need to consider providing coping/mitigating strategies—currently, the risk is not known, and such mitigating strategies are not routinely provided [27,28]. The risk/benefit of mood monitoring might be best investigated through an analysis of RCTs of frequent mood assessment, although other approaches, such as reports of adverse events, qualitative monitoring, and surveys from practice, all have their place in responsible technological innovation [29].

Previous reviews have explored mood monitoring but have not assessed efficacy in high-quality RCTs [28,30,31]. This is the first systematic review and meta-analysis that we are aware of that examines mood monitoring as an intervention in RCTs in BD. The aim of this systematic review is to assess the effect of mood tracking in people with BD and depression on relapse risk and symptoms of bipolar depression, mania/hypomania, and depression.

Methods

Overview

We used the Cochrane Handbook for Systematic Reviews of Interventions methodology and used a Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) checklist. The study was preregistered with the International Prospective Register of Systematic Reviews (PROSPERO: CRD42023396473 [32]).

Inclusion Criteria

The inclusion criteria were as follows: self-monitoring/ecological momentary assessment (EMA)/repeated symptom assessment in people with BD or depression as an intervention over a minimum period of 3 months, with rating of symptoms weekly at a minimum. On discussion with patient and public involvement, mood monitoring over periods less than 3 months might be misleading in various contexts associated with temporary change, such as changes in medication, menstrual, seasonal, or life event effects on mood, or as the person becomes familiar with a new method of mood monitoring. The studies needed to use an appropriate nonmood monitoring/EMA control. The studies should either use a validated measure of mood or validate the chosen measure with a validated mood measure. The studies could be published in any language and could be digital or nondigital, although we acknowledged that the majority of studies would use digital technologies. We only included RCTs with 20 or more participants with BD or depression [33]. We did not exclude RCT studies where mood monitoring was included but not the primary focus of the intervention (but it so happened that in the included studies, the mood monitoring did comprise a significant part of the overall intervention). We searched the gray literature (eg, conference abstracts, dissertations, policy literature, reports via ProQuest and Google Scholar—full details below) for unpublished studies that were eligible for inclusion.

Search Strategy and Selection Criteria

The complete search strategy is listed in the Multimedia Appendix 1. We searched PsychINFO, EMBASE, SCOPUS, IEEE Xplore, Ovid MEDLINE, Proquest SciTech Collection, ProQuest Dissertations and Theses Global, and Google Scholar using the search terms. Search results were exported for appraisal and stored on Rayyan [34]. The initial search was conducted on March 03, 2023, and updated on October 28, 2024. All abstracts were appraised by 2 independent screeners (LAW, GM, GS, RP, MM, and DP), and any disagreements were discussed, and a consensus arrived upon, with adjudication by a third independent screener if required. The full text of any potentially relevant papers was acquired, and if we were unable to source the full text of the study, we then contacted the corresponding author to request the paper. To determine if potentially relevant studies met the inclusion criteria, the full text was reviewed separately by 2 authors, again with discussion and consensus with a third reviewer if necessary. All papers for inclusion were reference checked along with relevant systematic reviews [15,18,27,28,30,31,35-38]. Key authors were also emailed to see if the inclusion of any ongoing unpublished studies could be included.

Data Extraction

Two independent reviewers extracted data (as per symptom severity scores for different time points) from studies meeting the inclusion criteria using identical data extraction forms. Irregularities in the data extraction were discussed, and any discrepancies were resolved through discussion.

Assessment of Study Bias

The Cochrane Collaboration’s Risk of Bias 2 tool was used for each study [39]. Risk of bias was assessed by 2 independent reviewers (LAW and GS), and any disagreement was resolved via discussion.

The certainty of the evidence for the meta-analysis results for each outcome was assessed independently and in duplicate by 2 review authors (LAW and GS) using the Grading of Recommendations, Assessment, Development, and Evaluation framework. This involved an individual assessment of each of the 5 domains of risk of bias (inconsistency, indirectness, imprecision, and publication bias), resulting in an overall assessment of the certainty of the evidence as “high,” “moderate,” “low,” or “very low” [40].

Synthesis of Results

The primary outcome in the meta-analysis for treatment studies was a reduction in depression/mania/hypomania incidence/symptoms for people with BD, and a reduction in depressive symptoms at 6 months postintervention. This timeframe was chosen to demonstrate the stability of treatment effects. For mania/hypomania/depression severity, we calculated standardized mean differences (SMDs). Insufficient studies examined BD or depression incidence or relapse risk to meta-analyze.

For outcomes included in more than one study, we measured statistical heterogeneity by calculating the I^2 statistic [41]. An I^2 of less than 30% was taken to indicate mild heterogeneity, and a fixed-effects model was used. When the I^2 was greater than or equal to 30%, a random-effects model was used. All analyses were performed using Review Manager (version 5.3; Cochrane Collaboration).

Results

Overview

The search identified a total of 23,515 studies. No studies that were not in English were found to meet the inclusion criteria. Following title and abstract screening, 21,638 studies were excluded, resulting in a total of 758 studies being reviewed in full. A total of 5 trials in people with BD and 3 trials in people with depression met the eligibility criteria and were included in the review. The BD trials included 803 participants, and the depression trials included 427 participants. Tables 2-4 display detailed characteristics of the studies and the mood monitoring protocols used. Figure 1 and Checklist 1 detail the search strategy with the PRISMA flowchart and checklist.

Table . Characteristics of included bipolar disorder studies.

Study	Country	Sample[a]	Age (years), mean (SD)	Female (%)[a]	Intervention	Comparator	Setting	Mood monitoring intervention	Active or passive mood monitoring	Mood monitoring duration (months)	Primary outcome
Faurholt-Jepsen et al [42] (n=67)	Denmark	Bipolar 1: 67%, Bipolar 2: 33%	29.3 (8.43)	67	MONARCA system plus: (1) Study nurse reviewing data and contacting patients if signs of deterioration to offer advice. (2) Self-monitored data graphically visualized.	Normal smartphone use	Secondary care: specialist mood disorder service for patients with a new diagnosis of bipolar or treatment resistance.	Daily smartphone self-monitoring: mood, sleep duration, medication taken, activity, irritability, mixed mood, cognitive problems, alcohol consumption, stress, menstruation, individualized EWS[b].	Active and passive	6	HDRS[c] and YMRS[d] at 1 - 6 months

Study	Country	Sample[a]	Age (years), mean (SD)	Female (%)[a]	Intervention	Comparator	Setting	Mood monitoring intervention	Active or passive mood monitoring	Mood monitoring duration (months)	Primary outcome
Faurholt-Jepsen et al [43] (n=129)	Denmark	Bipolar 1: 59%, Bipolar 2: 41%	43 (12)	59	Monsenso system plus: (1) Study nurse reviewing data and contacting patients if signs of deterioration to offer advice. (2) self-monitored data graphically visualized.	Normal smartphone use	Secondary care: specialist mood disorder service for patients with a new diagnosis of bipolar or treatment resistance	Daily smartphone self-monitoring items: mood, sleep duration, medication taken, activity, irritability, mixed mood, cognitive problems, alcohol consumption, stress, menstruation, individualized EWS, anxiety, self-defined personal parameters, free-text note. Objective smartphone data: phone usage, social activity, step count, GPS location.	Active and passive	9	HDRS and YMRS at 1 - 9 months

Study	Country	Sample[a]	Age (years), mean (SD)	Female (%)[a]	Intervention	Comparator	Setting	Mood monitoring intervention	Active or passive mood monitoring	Mood monitoring duration (months)	Primary outcome
Faurholt-Jepsen et al [44] (n=98)	Denmark	Bipolar 1: 58%, Bipolar 2: 42%	42.69 (13.46)	52	Monsenso system plus: (1) study nurse reviewing data and contacting patients if signs of deterioration to offer advice. (2) self-monitored data graphically visualized.	Usual care	Secondary care: specialist mood disorder service for patients with a new diagnosis of bipolar or treatment resistance.	Daily smartphone self-monitoring items: mood, sleep duration, medication taken, activity, irritability, mixed mood, cognitive problems, alcohol consumption, stress, menstruation, individualized EWS, anxiety, self-defined personal parameters, free-text note. Objective smartphone data: phone usage, social activity, step count, GPS location.	Active and passive	6	Rate and duration of psychiatric readmissions at 3 - 6 months

Study	Country	Sample[a]	Age (years), mean (SD)	Female (%)[a]	Intervention	Comparator	Setting	Mood monitoring intervention	Active or passive mood monitoring	Mood monitoring duration (months)	Primary outcome
Gliddon et al [45] (n=304)	Australia and the United States	Bipolar 1: 55%, Bipolar 2: 38%	39.47 (11.19)	82		Discussion forum	Mixed sample: participants recruited via advertising.	Online mood monitoring via MoodSwings and MoodSwings-Plus websites.	Active	12	MADRS[f] and YMRS at 3 - 12 months

Study	Country	Sample[a]	Age (years), mean (SD)	Female (%)[a]	Intervention	Comparator	Setting	Mood monitoring intervention	Active or passive mood monitoring	Mood monitoring duration (months)	Primary outcome
					Intervention 1: Discussion forum plus MoodSwings-Plus: MoodSwings plus additional CBT-based[e] interactive elements: tools to support mood and medication monitoring, life-chart development, cognitive strategies, motivational interviewing techniques, self-reflection, problem solving, identification of personal triggers and a relapse prevention plan. Intervention 2: Discussion forum plus MoodSwings: Online intervention comprising mood monitoring, assessing prodromal mood states, preventing relapse, and setting SMART goals. Online delivery of						

Study	Country	Sample[a]	Age (years), mean (SD)	Female (%)[a]	Intervention	Comparator	Setting	Mood monitoring intervention	Active or passive mood monitoring	Mood monitoring duration (months)	Primary outcome
					MAPS (Mood Assessment Prevent SMART) program.						
Goulding et al [14] (n=205)	The United States	Bipolar 1: 100%, Bipolar 2: 0%	42 (12)	61	Livewell	Usual care	Secondary care: 1 previous mood episode in the past year and current care by psychiatrist/nurse practitioner.	Smartphone-based self-management intervention: daily and weekly check-ins for weeks 1 - 16. Daily: adherence, sleep, duration, routine, wellness levels. Weekly: symptom severity scoring for all individual DSM-IV[g] mood symptoms.	Active	4	Time to relapse

[a]Reference articles do not provide the complete statistics, so absolute values corresponding to percentage values cannot be provided here.

[b]EWS: Early Warning Signs.

[c]HDRS: Hamilton Depression Rating Scale 17 items.

[d]YMRS: Young Mania Rating Scale.

[e]CBT: cognitive behavior therapy.

[f]MADRS: Montgomery-Åsberg Depression Rating Scale.

[g]DSM: Diagnostic and Statistical Manual of Mental Disorders.

Table . Intervention protocols of included studies for depression and bipolar disorder.

Study type and respective included studies	Mood monitoring intervention	Mood monitoring duration (months)	Active or passive mood monitoring	Analog or digital mood monitoring	Adherence to mood monitoring	Trial attrition
Included studies for bipolar disorder						
Faurholt-Jepsen et al [42]	Daily smartphone self-monitoring: mood, sleep duration, medication taken, activity, irritability, mixed mood, cognitive problems, alcohol consumption, stress, menstruation, individualized EWS[a], clinical feedback loop.	6	Active and passive	Digital	>93% of patients randomized to the intervention group self-reported on a daily basis.	Intervention group: 3% attrition over 6 months, control group: 3% attrition over 6 months.
Faurholt-Jepsen et al [43]	Daily smartphone self-monitoring items: mood, sleep duration, medication taken, activity, irritability, mixed mood, cognitive problems, alcohol consumption, stress, menstruation, individualized EWS, anxiety, self-defined personal parameters, free-text note. Objective smartphone data: phone usage, social activity, step count, GPS location, clinical feedback loop.	9	Active and passive	Digital	Over 9 months, patients in the intervention group adhered to the daily self-monitoring 72.6% of the days.	Intervention group: 7% attrition at 9 months, control group: 7% attrition at 9 months.
Faurholt-Jepsen et al [44]	Daily smartphone self-monitoring items: mood, sleep duration, medication taken, activity, irritability, mixed mood, cognitive problems, alcohol consumption, stress, menstruation, individualized EWS, anxiety, self-defined personal parameters, free-text note. Objective smartphone data: phone usage, social activity, step count, GPS location, clinical feedback loop.	6	Active and passive	Digital	80.6% adherence to daily self-monitoring in intervention group over 6 months.	Total attrition: 35% at 6 months, intervention group: 22%, control group: 53%.

Study type and respective included studies	Mood monitoring intervention	Mood monitoring duration (months)	Active or passive mood monitoring	Analog or digital mood monitoring	Adherence to mood monitoring	Trial attrition
Gliddon et al [45]	Intervention 1: Discussion forum plus MoodSwings-Plus: MoodSwings plus additional CBT-based[b] interactive elements—tools to support mood and medication monitoring, life-chart development, cognitive strategies, motivational interviewing techniques, self-reflection, problem solving, identification of personal triggers, and a relapse prevention plan. Intervention 2: Discussion forum plus MoodSwings: Online intervention comprising mood monitoring, assessing prodromal mood states, preventing relapse, and setting SMART goals. Online delivery of MAPS (Mood Assessment Prevent SMART) program.	12	Active	Digital	Control group: 89% accessed the discussion forum, MoodSwings group: 86% accessed the modules, MoodSwings-Plus: 74% accessed the tools.	Total attrition: 9% at 12 months, control group: 6%, MoodSwings group: 7%, MoodSwings-Plus: 13%.
Goulding et al [14]	Smartphone-based self-management intervention: daily and weekly check-ins for weeks 1 - 16. Daily: adherence, sleep, duration, routine, wellness levels. Weekly: symptom severity scoring for all individual DSM-IV[c] mood symptoms.	4	Active	Digital	The mean (SE) percentage of daily check-ins completed during weeks 1 through 4 was 78% (3%), 74% (3%), 71% (3%), and 64% (3%), respectively, 66% (3%) during week 6, and 47% (4%) during week 16.	Intervention group: 15% attrition at 4 months, control group: 15% attrition at 4 months.
Included studies for depression						
Aikens et al [46]	Automated Interactive Voice Response telephone calls assessing symptom severity: PHQ-9[d] and antidepressant adherence.	12	Active	Analog—telephone	22 % in intervention arm completed<50% of scheduled calls	Total attrition: 14%, intervention: 17%, control: 10%.

Study type and respective included studies	Mood monitoring intervention	Mood monitoring duration (months)	Active or passive mood monitoring	Analog or digital mood monitoring	Adherence to mood monitoring	Trial attrition
Tonning et al [47]	Monsenso system plus: (1) Study nurse reviewing data and contacting patients if signs of deterioration to offer advice, (2) self-monitored data graphically visualized, and (3) smartphone-based CBT modules	6	Active and passive	Digital	82.7% in intervention arm	Total attrition: 82.5%, intervention: 20%, control: 15%.
Hunkeler et al [48]	Personalized self-monitoring via eCare for Moods—tracking health-related disability, medication adherence, side effects, alcohol and drug use, new symptoms, early warning signs. Graphs of monitoring data displayed over time.	24	Active	Digital	≈87% entered any monitoring data over the first 6 months, ≈45 % entered any monitoring data over the second 6 months	Total attrition: 16%, intervention attrition: 22%, control attrition: 12%.

[a]EWS: early warning signs.

[b]CBT: cognitive behavior therapy.

[c]DSM: Diagnostic and Statistical Manual of Mental Disorders.

[d]PHQ-9: Patient Health Questionnaire-9.

Table . Characteristics of included depression studies.

Study	Country	n	Age (years), mean (SD)	Female (%)[a]	Intervention	Comparator	Setting	Mood monitoring intervention	Active or passive mood monitoring	Mood monitoring duration	Primary outcome
Aikens et al [46]	The United States	204	49[b]	81	Automated Interactive Voice Response telephone calls	Enhanced usual care: usual care plus printed self-management material at baseline and assigned family/friend to discuss this with weekly	Primary care	Automated Interactive Voice Response telephone calls assessing symptom severity—PHQ-9[c] and antidepressant adherence.	Active	12 months	PHQ-9 at 6 - 12 months
Tonning et al [47]	Denmark	120	Intervention: 44.5 (14.0), Control: 43.4 (14.3)	Intervention: 47.5 (28), Control: 43.4 (14.3)	Monsenso system plus: 1. study nurse reviewing data and contacting patients if sign of deterioration to offer advice 2. self-monitored data graphically visualized 3. smartphone-based CBT[d] modules	TAU[e]	Tertiary care: specialist mood disorder service for patients with a new diagnosis of bipolar or treatment resistance	Daily smartphone self-monitoring items: mood, sleep duration, medication taken, activity, irritability, mixed mood, cognitive problems, alcohol consumption, stress, menstruation, individualized EWS[f], anxiety, self-defined personal parameters, free-text note. Objective smartphone data: phone usage, social activity, step count, GPS location	Active and passive	6 months	Rate and accumulated duration of psychiatric readmissions. (HDRS-17[g] as secondary outcome)

Study	Country	n	Age (years), mean (SD)	Female (%)[a]	Intervention	Comparator	Setting	Mood monitoring intervention	Active or passive mood monitoring	Mood monitoring duration	Primary outcome
Hunkeler et al [48]	The United States	103	Intervention: 48.49 (12.83), Usual care: 51.88 (10.56)	79.6	eCare for moods: website offering personalized self-monitoring, messaging with eCare manager, depression psychoeducation, CBT modules, online discussion group, problem-specific advice, personal database, task lists, appointment calendar.	TAU	Secondary care	Personalized self-monitoring via eCare for moods: tracking health-related disability, medication adherence, side effects, alcohol and drug use, new symptoms, early warning signs. Graphs of monitoring data displayed over time.	Active	12 months	Psychiatric Status Rating for Depression of 6 questions adapted from SCID[h] measured weekly

[a]Reference articles do not provide the complete statistics, so absolute values corresponding to percentage values cannot be provided here.

[b]Reference articles do not provide the complete statistics, so the SD value cannot be provided alongside mean value cannot be provided here.

[c]PHQ-9: Patient Health Questionnaire-9.

[d]CBT: cognitive behavior therapy.

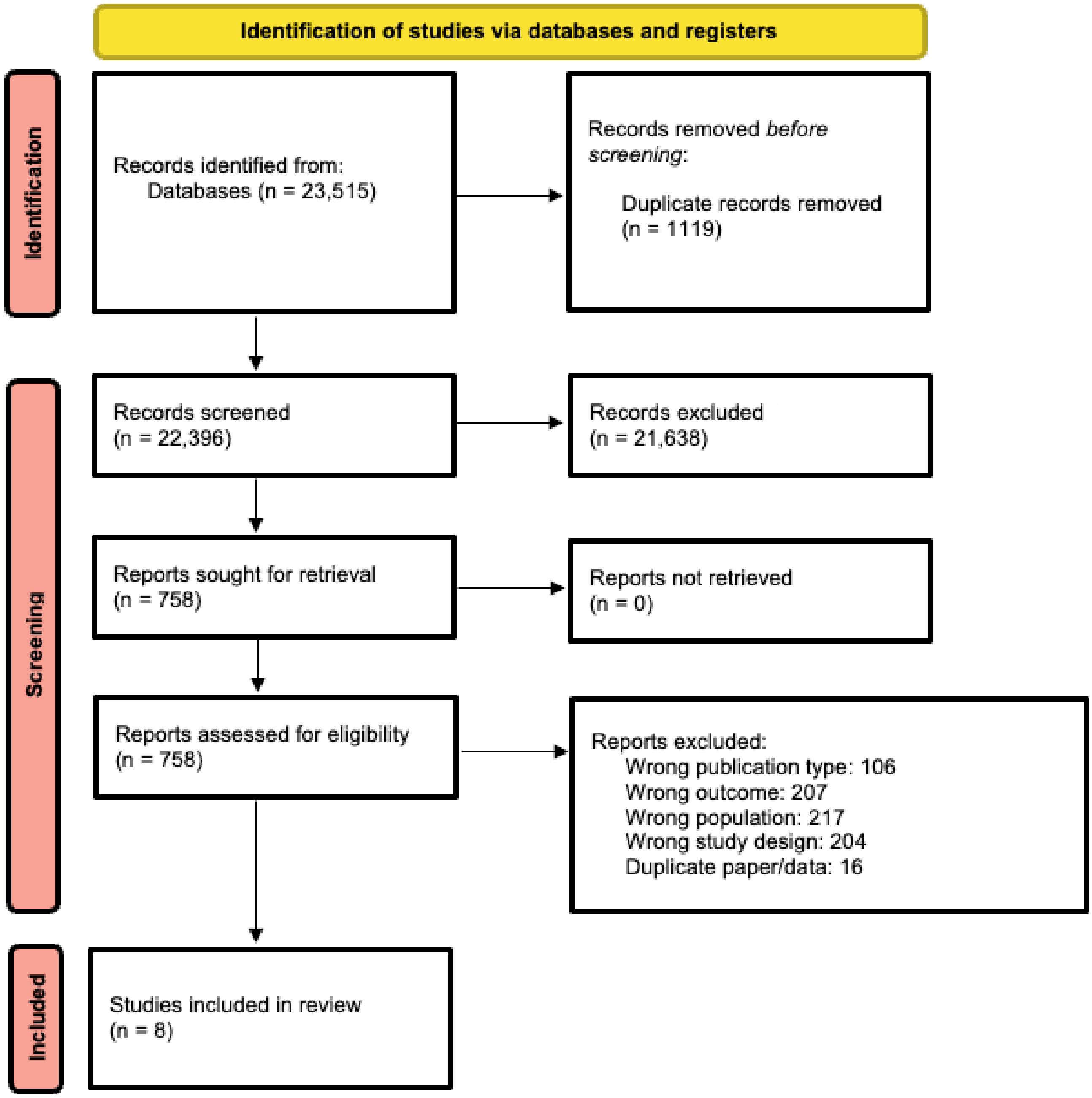
[e]TAU: treatment as usual.

[f]EWS: Early Warning Signs.

[g]HDRS-17: Hamilton Depression Rating Scale 17 items.

[h]SCID: Structured Clinical Interview for DSM-IV.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flowchart.



Bipolar Disorder: Overview of Individual Study Findings

In BD, 5 trials [14,43-45,49-61] used 3 different mood monitoring protocols. These were the MONARCA/Monsenso system, Livewell, and MoodSwings [45,52]. Only the MONARCA/Monsenso system incorporated passive data capture into the intervention. These were all digital; 3 used active and passive ambulatory assessment [43,44,51] while 2 just used active ambulatory assessment [14,45].

All studies assessed mania and depression severity using a mixture of self-report and clinical rating scales, with three of these as the primary outcome [43,45,62]. Two studies assessed relapse rate/psychiatric readmission or duration as the primary outcome [44]. Only one study [45] provided raw mania/depression severity data, and we contacted the authors of the other papers to obtain this for the analyses. Studies used similar inclusion criteria, but there were some key differences. All studies recruited individuals with BD from clinical services and confirmed BD via a structured clinical interview at baseline and then relied on self-report measures/clinical ratings for outcome assessment. All studies excluded individuals who were currently experiencing a major mood episode, either by using a cutoff on self-report scores/clinical rating scales (which varied between studies) [14,42,45] or by the participant completing treatment at a specialist mood disorder service [43], with one study recruiting individuals on discharge from inpatient care following hospitalization for an affective episode [44].

The results of the trials were mixed. The 2 trials [14,47] assessing relapse/readmission risk showed no effect of mood monitoring. Goulding et al [14] demonstrated an effect of decreased relapse risk for low-risk individuals, but no effect on percentage-time symptomatic for all participants. Three studies did not identify a decrease in depression or mania scores from mood monitoring on clinical ratings conducted blinded to allocation [43,51,63]. Gliddon et al [45] reported decreases in depressive symptoms when compared with the peer support control, Goulding et al [14] also reported improvements in depression severity alongside improved relational quality of life.

Faurholt-Jepsen et al [51] and Faurholt-Jepsen et al [43] found a nonstatistically significant trend for worsening depressive symptoms versus the control group.

Bipolar Disorder: Meta-Analysis

Concerning the primary outcome, there was no effect of mood monitoring interventions in reducing symptoms of mania/hypomania (Figure 2: 6 comparisons, n=873; SMD −0.16, 95% CI −0.34 to 0.01; P=.06; I^2=36%) or bipolar depression (Figure 3: 6 comparisons, n=873; SMD −0.08, 95% CI −0.31 to 0.15; P=.02; I^2=63%).

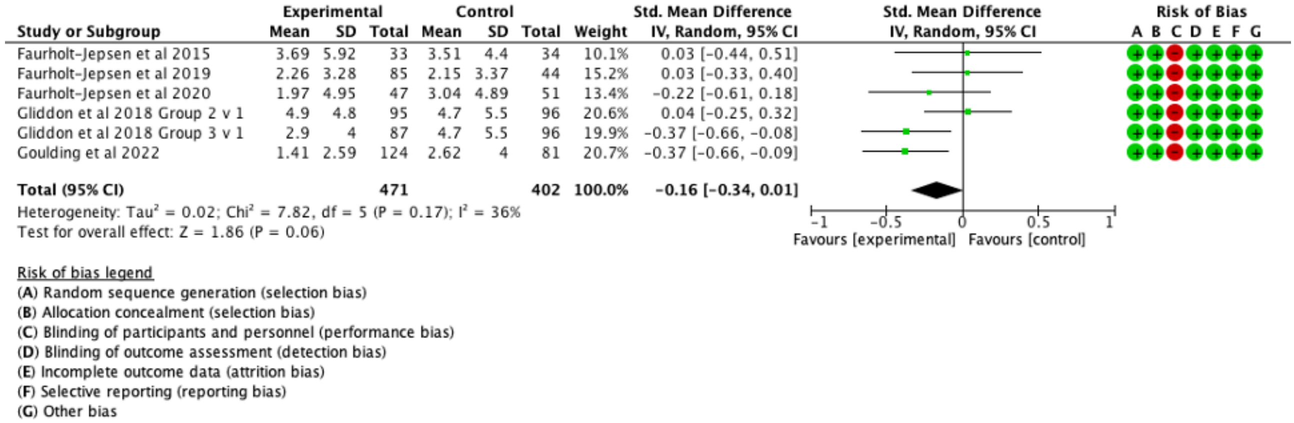


Figure 2. Forest plot of effects of mood monitoring interventions for the treatment of symptoms of mania/hypomania in people with BD at 6 - 12 months [14,42-45]. BD: bipolar disorder.

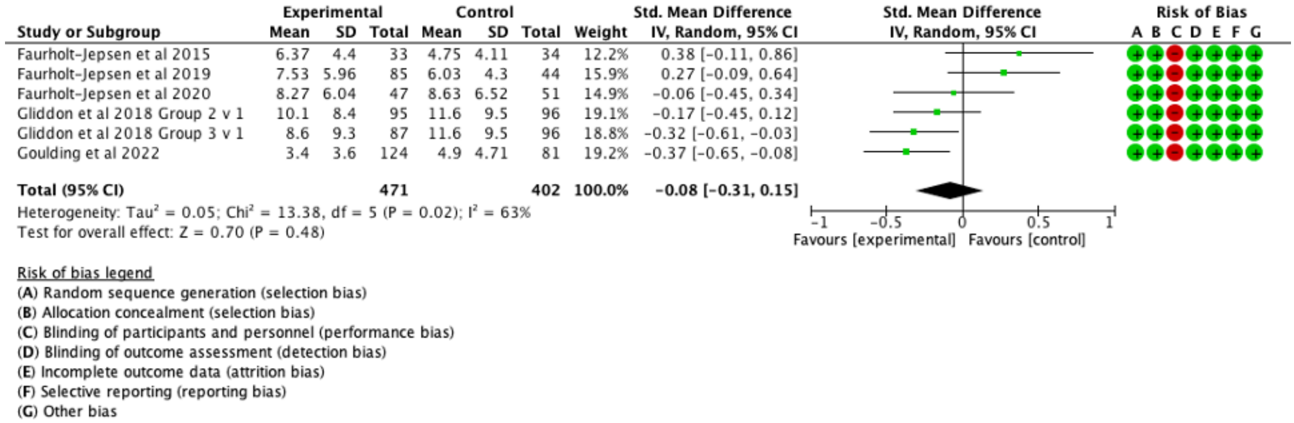


Figure 3. Forest plot of effects of mood monitoring interventions for the treatment of symptoms of depression in people with BD at 6 - 12 months [14,42-45].

Depression: Overview of Individual Study Findings

In depression, 3 trials used 3 different mood tracking procedures. These were: interactive voice response telephone calls (IVR) [46], Monsenso [47], and eCare for Moods [48]. None of the trials used mood monitoring as the primary outcome; instead, using standardized infrequent assessments of mood (Tables 2 and 4). One study was analog using the telephone [46], the other two were digital [47,48]; two used active ambulatory assessment [48], with one using active and passive ambulatory assessment [47].

The Monsenso [47] and eCare for Moods [48] protocols used daily monitoring, while IVR calls [46] were administered weekly. Only the Monsenso system incorporated passive data collection. All protocols incorporated clinical feedback of the assessment as an intervention. Monseno and eCare for Moods used digital technology, while IVR used telephone calls and voice messages. We did not identify any paper-based charting trials. The duration of ambulatory assessment protocols varied from 6 months to 24 months.

Two studies assessed relapse rate/psychiatric readmission or duration as the primary outcome [46,48]. Tonning et al [64] assessed the rate and accumulated duration of psychiatric readmissions as the primary outcome, assessing depression severity as a secondary outcome. Only one study [48] provided raw depression severity data, and we contacted the authors of the other papers to obtain this for the analyses. Studies used similar inclusion criteria, but there were some key differences. All studies recruited individuals with depression from clinical services. Two studies confirmed depression via a structured clinical interview at baseline, while Aikens et al [46] confirmed depression via the Patient Health Questionnaire-9. All studies then relied on self-report measures for outcome assessment. Aikens et al [46] excluded patients who were experiencing major psychiatric distress and recruited from community samples. Hunkeler et al [48] did not exclude individuals with suicidal ideation or a particular severity of depression, again recruiting from community clinics. Tonning et al [47] recruited people with depression receiving inpatient care, providing the intervention postdischarge.

The results of the 3 trials were mixed. Tonning et al [47] report no change in relapse risk or readmission duration, as well as no change in depressive symptoms. They did, however, report a range of benefits across tertiary outcomes when adjusted for age, sex, and Hamilton Depression Rating Scale scores. Patients in the intervention group reported statistically higher recovery, measured using the Recovery Assessment Scale, as well as a tendency (not statistically significant) towards higher quality of life, higher well-being, more satisfaction with treatment, and higher behavioral activation in the intervention group compared with the control group.

In eCareformoods, participants in the intervention group experienced a statistically significant reduction in depressive symptoms at 2 years. A higher proportion of those in the intervention group remained in recovery from their depression, and the number needed to treat was calculated at 8. Intervention participants also had improvements across a range of secondary outcomes, including improved general mental health, learning new coping skills, greater satisfaction with specialty care, and more confidence in managing depression. These were all statistically significant.

Aikens et al [46] report a statistically significant improvement in Patient Health Questionnaire-9 depression severity (2.5 points) at 6 months in the intervention group. This persisted for 12 months. Clinical response was more likely in the intervention group than the control at 6 months, but this difference decreased in size and lost statistical significance by 12 months.

Only Tonning et al [47] reported adverse effects, and these were as follows: 3 participants found the monitoring stressful, and 1 participant did not find it helpful.

Unipolar Depression: Meta-Analysis

There was a small effect of borderline statistical significance at 12 months (Figure 4: 2 comparisons, n=262; SMD −0.25, 95% CI −0.49 to 0.00; P=.05; I^2=45%) but not at 6 months (Figure 5: 2 comparisons, n=268; SMD −0.21, 95% CI −0.54 to 0.12; P=.21; I^2=12%). Only 2 trials were included in the meta-analysis as the appropriate data for Tonning et al [47] were not available.

Figure 4. Forest plot of effects of mood monitoring interventions in reducing symptoms of depression at 12 months [46,48].

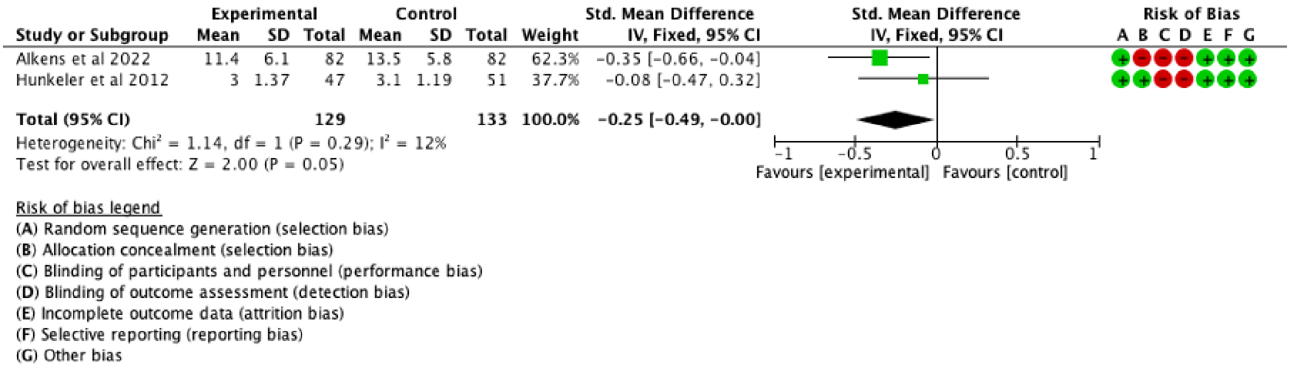
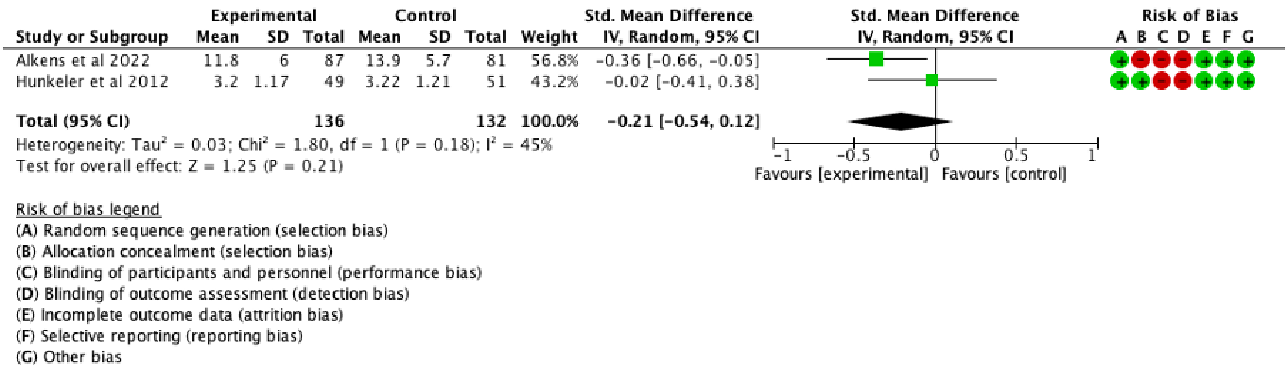


Figure 5. Forest plot of effects of mood monitoring interventions in reducing symptoms of depression at 6 months [46,48].



Risk of Bias Assessments

The quality of these RCTs was good, with all trials having low risk of bias (Table 5). All trials used intention-to-treat analysis. All studies reported adherence and attrition.

Table . Risk of bias assessments for included bipolar disorder and depression studies.

Study	Risk of bias criteria							
	Random sequence generation	Allocation concealment	Blinding of participants and personnel	Blinding of outcome assessment	Incomplete outcome data	Selective reporting	Other sources of bias	Total number of low-risk domains
Risk of bias assessments for included bipolar disorder trials								
Faurholt-Jepsen et al [42]	Low risk	Low risk	High risk	Unclear	Low risk	Low risk	Low risk	6
Faurholt-Jepsen et al [43]	Low risk	Low risk	High risk	Low risk	Low risk	Low risk	Low risk	6
Faurholt-Jepsen et al [44]	Low risk	Low risk	High risk	Low risk	Low risk	Low risk	Low risk	6
Gliddon et al [45]	Low risk	Low risk	High risk	Low risk	Low risk	Low risk	Low risk	6
Goulding et al [14]	Low risk	Low risk	High risk	Low risk	Low risk	Low risk	Low risk	6
Risk of bias assessments for included depression trials								
Aikens et al [46]	Low risk	Low risk	High risk	High risk	Low risk	Low risk	Low risk	5
Tonning et al [47]	Low risk	Low risk	High risk	Low risk	Low risk	Low risk	Low risk	6
Hunkeler et al [48]	Low risk	Low risk	High risk	High risk	Low risk	Low risk	Low risk	5

Discussion

Principal Findings

This systematic review of mood monitoring interventions in people with BD found no effect either way on symptoms of mania/hypomania or bipolar depression in people with BD at 6 - 12 months. There was no robust evidence of mood monitoring either increasing or decreasing symptoms of depression, with no effect at 6 months and borderline statistical improvement at 12 months in only 2 RCTs. There were some other benefits of mood monitoring across 3 RCTs in depression, but there was no consistency in what was measured or the outcomes that were improved.

Bipolar Disorder

Mood monitoring is theorized to work by improving understanding and insight to enable people to self-manage their BD [16]. People with BD seem to use the data provided to them by mood monitoring in varied ways [65]. These appear to be highly personalized and tailored to what works best for them. The mood monitoring protocol provides a platform for people to interpret their own mood data, devising highly personal ways of self-managing their BD subsequently [66]. This self-management may focus on sleep, medication, crisis planning, and communication in close relationships [16]. Thus, the variability in these outcomes might reflect the different ways in which participants used the mood monitoring information and their coping strategies in the face of depression and mania [8] as well as differences in populations and measurement of mood. Despite the equivocal results reported here, the practice remains hugely popular, and in the digital sphere, there are multiple different apps and protocols available aimed at people with BD, such as the Bipolar UK mood tracker [4], eMoods [42], Moodily, Moodnotes, etc [3].

Depression

There is some agreement with the findings from the meta-analysis we report from RCTs of mood monitoring in bipolar depression that, on the whole, mood monitoring has neutral effects on outcome, but there is quite a lot of heterogeneity, with some people reporting benefits, and others distress and burden. The findings are not robust. While all 3 RCTs report other benefits of mood monitoring, such as relapse, recovery, increased overall mental health, greater confidence in managing depression, and greater satisfaction with mental health services, there was no consistency in what or how these secondary or tertiary outcomes were measured. Only one of these RCTs reported adverse effects of mood monitoring, with 3 participants finding it a burden and stressful. From qualitative research, some variables that have bearing on the outcome of mood monitoring are context in which mood monitoring is taking place, the usual coping strategies, the nature of mood monitoring, and the population that is being examined. Only one RCT used passive mood monitoring, and none used EMA approaches, so there is no evidence available on modern developments in mood monitoring of depression symptoms in people with unipolar depression. Despite the modest effects reported here, mood monitoring remains popular for people with depression [67-70], and in the digital sphere, there are multiple different apps/protocols available [70]. This review highlights that the popularity of the process may be disproportionate to the direct effects of mood monitoring as an intervention. From the systematic reviews of qualitative data that we have performed, and like self-monitoring of BD, the popularity of mood monitoring in depression may be through the empowerment of the individual to use the data from mood monitoring in a variety of personal ways rather than any direct effect on any outcome [21,22]. However, mood monitoring is not for everyone; for some, it is a burden or a reminder of poor well-being.

Strengths and Limitations

Our search was thorough and in accordance with Cochrane methodology. We also consulted experts in the field, used a wide search, and used reference searching. Our conclusions, however, were limited due to a paucity of literature. This itself is an important finding, considering the possible implications of frequent mood assessment as ambulatory assessment and EMA approaches develop further. Sometimes there is definitional overlap between EMA protocols and mood monitoring protocols [17]. We did not identify a large enough group of studies using EMA approaches, passive monitoring, or a combination of the two to determine any outcomes or harms from these approaches in this review. We did not include mood monitoring protocols that used shorter follow-up periods, as some previous reviews have done [18-20], because we wanted to examine whether there was any high-quality evidence on benefits or harms. Therefore, we chose RCTs of mood monitoring interventions lasting at least 3 months and clinically relevant follow-up lengths of at least 6 - 12 months.

Many other trials were excluded because control groups also had an intervention consisting of some element of mood monitoring. In all the included studies, the main therapeutic intervention applied to all participants in the intervention group was mood monitoring, but all of these trials included additional likely therapeutic elements, which makes isolating the effect of the mood monitoring challenging. With regard to the 5 BD trials, the MONARCA/Monsenseo [43,44,51] studies included data review and outreach by a nurse; the Livewell [14] and Moodswings [45] trials included additional coaching, psychoeducational materials, and planning for management of early warning signs and symptoms. In all 3 RCTs of mood monitoring in depression, there were other elements that might have improved or worsened depression symptoms. Two RCTs used cognitive behavior therapy modules [47,48], and one provided self-management guidance based on the severity of the participant's symptoms [46]. These additional elements may prove therapeutic benefit but also obfuscate any effect directly from mood monitoring itself. Thus, it may be impossible to determine only the effects of mood monitoring on symptoms of depression or mania because interventions have elements of other interventions as well. There were no studies with just a basic mood monitoring element versus a nonmood monitoring control.

In addition to nonmood monitoring elements of the intervention, there may be factors other than the differences in the intervention that explain some of the mixed results we observe here. These include clinical characteristics of the sample such as the type/subtype of mood disorder (Bipolar 1: Bipolar 2 proportions reported in Tables 2 and 4) and as the results of Goulding et al [14] suggest, there may be improved beneficial effects in lower-risk individuals with BD who are at a specific stage of their illness, and this is supported by qualitative work suggesting that there may be a right time for people with BD to be using mood tracking as an intervention [22,71]. Furthermore, while we did not assess adherence to mood tracking in this paper, we have addressed this in a separate meta-analysis [72], and this separate work demonstrates that all of the studies included here had >70% adherence. Thus, it is unclear if poor adherence levels may cause a failure of effect of mood tracking. Suboptimal adherence may obfuscate these effects, particularly if the effect size is small. However, poor adherence is pragmatic and reflective of real-life outcomes [73], and so these results provide a more pragmatic signal of effect.

Our review focused on the effects of mood monitoring on mood symptoms. However, there might be other benefits from mood monitoring, such as the development of mental literacy about the condition early in its course by seeing how mood varies in severity across time or learning to recognize the symptoms of mania. It may improve confidence or coping strategies that exert some control over the symptoms through developing insights into the conditions. These elements may improve recovery and function [74]. For instance, a valuable role for mood monitoring might be to help with decision-making when seeking help for those who frequently relapse with depression or BD, and when to make important life decisions, such as new responsibilities or decisions with an element of risk, like taking a holiday abroad, away from usual sources of help. Thus, mood monitoring might also have different clinically important outcomes depending on the recency of diagnosis, course of illness, or where recovery and improvement of function are a clinical priority.

Future Research and Clinical Implications

Future research in mood disorders should evaluate the definitive efficacy of mood monitoring alone rather than additional components to assess for benefits and potential harms. Care should be taken over control groups to ensure that they do not unwittingly include mood monitoring or psychotherapeutic approaches that might obscure the effects of mood monitoring. It remains unclear for whom mood monitoring may be most effective, and future research should assess this. It could be, for example, that it is least effective and most harmful in people with mood disorders who cope with depression by suppressing these symptoms and have not developed other coping strategies. Other groups who may experience harm might be people with mood disorders that feature paranoia or where monitoring of the person was used as a form of coercion, for example, as a feature of morbid jealousy or other abusive relationships. The benefits and harms from more modern approaches to mood monitoring should be explored with a broader range of outcomes focusing on mood monitoring to improve recovery, function, capability, and quality of life.

The field would benefit from a definitive RCT assessing time to relapse in those in asymptomatic remission with an appropriate control group. This would use a mood monitoring intervention with minimal additional psychotherapeutic strategies. This work is important due to the increased use of ambulatory assessment measures as measures of treatment outcomes [75] in studies looking at other interventions and not primarily to explore the effects of measurement itself. These are, in many cases, indistinguishable from mood monitoring protocols that study mood monitoring as an intervention. Such a RCT of mood monitoring might benefit from qualitative work performed alongside to better understand how participants use information gleaned from mood monitoring, as they are not always passive consumers of such information [16]. Finally, we also need to examine mood monitoring in poorer and ethnically diverse populations. These populations may be excluded from digital interventions through poverty or other disadvantages [76], and the characteristics that predispose a population to digital disadvantage are the same as those that might put them at increased risk of mental illness [77,78]. If digital forms of monitoring and interventions are to be used more broadly in health care, they may have the effect of widening pre-existing health inequalities through lack of access to the technologies themselves, as well as to research, in disadvantaged populations. Technology could allow populations who might not otherwise easily access health care to access more relevant information or interventions, in languages other than the one used by the health care provider.

There is insufficient robust, high-quality evidence of benefits or harm to recommend active mood monitoring by participants as primary outcomes in an RCT. In fact, the heterogeneity of outcomes and variability in the way people with mood disorders appraise and use active mood monitoring suggests that active mood monitoring would be unsuitable for use as a primary outcome; a source of variance in outcome might be introduced that would be nonrandom and not necessarily predictable. Since many passive mood monitoring and ambulatory assessments are more implicit measures of mood, participants may not have as much agency in appraising and using such information. However, at this point, these measures are also not suitable as their effects on both outcomes and harms have not been sufficiently tested.

The meta-analysis showed small, nonsignificant, or borderline effects with the use of mood-monitoring, and understanding the practical implications of this is important, particularly considering the popularity of mood-tracking apps, some of which carry recommendations by organizations such as Bipolar UK. Mood monitoring, often coupled with other additional elements in the forms assessed here, did not have clinically important effects, and the qualitative research frequently reporting benefit does not align with the quantitative evidence presented here [21,79]. It is difficult to know under what circumstances mood monitoring may work for these individuals who do report a qualitative benefit. The qualitative research suggests that just mood monitoring is insufficient for any clinical effect and that individuals must incorporate this into a wide range of self-management strategies to keep them well [21,22,79]. The finding that there is no large clinical or placebo effect from mood monitoring suggests that these tools may make excellent outcome measures for research and clinical practice—fundamentally monitoring symptoms over time and using this information as an adjunct to make decisions around care (eg, improvements in shared decision making using more accurate data) and improving existing research outcomes.

Conclusions

As technological advances are applied to digital health and the capacity for more usable, passive mood monitoring increases, it is vital to understand whether these interventions work and for whom. It is also important to understand any positive or negative effects, as self-tracking is often used as a control or outcome assessment method in studies [17]. For BD, this review showed no effect of mood monitoring on symptoms of mania or bipolar depression, although the evidence was not robust with moderate to high heterogeneity in outcome. In people with depression, there was no robust evidence of the effects of mood monitoring on depression symptoms, with only 2 RCTs contributing to the meta-analysis. There was no evidence of the effect of mood monitoring using EMA. These results initially suggest that ambulatory assessment does not induce large placebo effects or significantly negatively/positively affect mood, and thus that mood monitoring may be an appropriate outcome measure for research or for clinical practice.

Acknowledgments

We would like to thank the University of Nottingham Library Service for their assistance with formulating the search terms. RM was chief investigator on a grant from the UK National Institute for Health and Care Research (NIHR) Applied Research Collaboration (ARC) East Midlands Mental Health and Well-being Theme. RM has received other NIHR funding for research on interventions for depression and bipolar disorder, including the lead for the Nottingham NIHR Biomedical Research Center mental health and technology theme and the NIHR MindTech Health Technology Collaborative.

Funding

This work was supported by a Wellcome Trust Fellowship to LAW and MM.

Authors' Contributions

LAW and RM were responsible for the original study design. LAW, GS, DP, and MM were responsible for the search, identification of relevant papers, data extraction, and risk of bias assessments. LAW was responsible for data analysis and initial drafting of the report. LAW and RM were responsible for the subsequent interpretation, editing, and rewriting of the report.

Conflicts of Interest

RM was chief investigator on a grant from the UK National Institute for Health and Care Research (NIHR) Applied Research Collaboration (ARC) East Midlands. RM has received other NIHR funding for research on interventions for depression and has received funding from Novartis to serve on a data management and ethics committee for 2 trials on the treatment of depression.

Multimedia Appendix 1
Search strategy.
[DOCX File, 26 KB - mental_v13i1e84020_app1.docx]

Checklist 1
PRISMA checklist.
[DOCX File, 18 KB - mental_v13i1e84020_app2.docx]

References

1. Majid S, Morriss R, Figueredo G, Reeves S. Exploring self-tracking practices for those with lived experience of bipolar disorder: learning from combined principles of patient and public involvement and HCI. DIS ’22: Proceedings of the 2022 ACM Designing Interactive Systems Conference 2022 Jun 13. [doi: 10.1145/3532106.3533531]
2. Majid S, Reeves S, Figueredo G, et al. The extent of user involvement in the design of self-tracking technology for bipolar disorder: literature review. JMIR Ment Health 2021 Dec 20;8(12):e27991. [doi: 10.2196/27991] [Medline: 34931992]
3. Hope B. Top 7 mood-tracking apps for bipolar disorder. bphope. 2024. URL: https://www.bphope.com/bipolar-buzz/bipolar-lifestyle-top-6-apps-to-track-your-moods/ [accessed 2025-9-17]
4. Track your mood. Bipolar UK. URL: https://www.bipolaruk.org/get-support/track-your-mood/ [accessed 2025-08-17]
5. de Azevedo Cardoso T, Kochhar S, Torous J, Morton E. Digital tools to facilitate the detection and treatment of bipolar disorder: key developments and future directions. JMIR Ment Health 2024;11:e58631. [doi: 10.2196/58631]
6. Morton E, Nicholas J, Yang L, et al. Evaluating the quality, safety, and functionality of commonly used smartphone apps for bipolar disorder mood and sleep self-management. Int J Bipolar Disord 2022 Apr 4;10(1):10. [doi: 10.1186/s40345-022-00256-6] [Medline: 35368207]
7. Nicholas J, Larsen ME, Proudfoot J, Christensen H. Mobile apps for bipolar disorder: a systematic review of features and content quality. J Med Internet Res 2015 Aug 17;17(8):e198. [doi: 10.2196/jmir.4581] [Medline: 26283290]
8. Mood Tracker app. Bipolar UK. 2025. URL: https://www.bipolaruk.org/get-support/track-your-mood/mood-tracker-app/ [accessed 2025-08-17]
9. Lagan S, D’Mello R, Vaidyam A, Bilden R, Torous J. Assessing mental health apps marketplaces with objective metrics from 29,190 data points from 278 apps. Acta Psychiatr Scand 2021 Aug;144(2):201-210. [doi: 10.1111/acps.13306] [Medline: 33835483]
10. Qu C, Sas C, Daudén Roquet C, Doherty G. Functionality of top-rated mobile apps for depression: systematic search and evaluation. JMIR Ment Health 2020 Jan 24;7(1):e15321. [doi: 10.2196/15321] [Medline: 32012079]
11. Schueller SM, Neary M, Lai J, Epstein DA. Understanding people’s use of and perspectives on mood-tracking apps: interview study. JMIR Ment Health 2021;8(8):e29368. [doi: 10.2196/29368]
12. Caldeira C, Chen Y, Chan L, Pham V, Chen Y, Zheng K. Mobile apps for mood tracking: an analysis of features and user reviews. AMIA Annu Symp Proc 2017;2017:495-504. [Medline: 29854114]
13. Bai R, Xiao L, Guo Y, et al. Tracking and monitoring mood stability of patients with major depressive disorder by machine learning models using passive digital data: prospective naturalistic multicenter study. JMIR Mhealth Uhealth 2021;9(3):e24365. [doi: 10.2196/24365]
14. Goulding EH, Dopke CA, Rossom R, Jonathan G, Mohr D, Kwasny MJ. Effects of a smartphone-based self-management intervention for individuals with bipolar disorder on relapse, symptom burden, and quality of life: a randomized clinical trial. JAMA Psychiatry 2023 Feb 1;80(2):109-118. [doi: 10.1001/jamapsychiatry.2022.4304] [Medline: 36542401]
15. Koenders MA, Nolen WA, Giltay EJ, Hoencamp E, Spijker AT. The use of the prospective NIMH Life Chart Method as a bipolar mood assessment method in research: a systematic review of different methods, outcome measures and interpretations. J Affect Disord 2015 Apr 1;175:260-268. [doi: 10.1016/j.jad.2015.01.005] [Medline: 25658502]
16. Astill Wright L, Moore M, Reeves S, Vallejos EP, Morriss R. Improving the utility, safety, and ethical use of a passive mood-tracking app for people with bipolar disorder using coproduction: qualitative focus group study. JMIR Form Res 2025 Feb 7;9:e65140. [doi: 10.2196/65140] [Medline: 39918865]
17. De Angel V, Lewis S, White K, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ Digit Med 2022 Jan 11;5(1):3. [doi: 10.1038/s41746-021-00548-8] [Medline: 35017634]
18. Wenze SJ, Miller IW. Use of ecological momentary assessment in mood disorders research. Clin Psychol Rev 2010 Aug;30(6):794-804. [doi: 10.1016/j.cpr.2010.06.007] [Medline: 20619520]

19. Morriss R, Vinjamuri I, Faizal MA, Bolton CA, McCarthy JP. Training to recognise the early signs of recurrence in schizophrenia. Cochrane Database Syst Rev 2013 Feb 28;2013(2):CD005147. [doi: 10.1002/14651858.CD005147.pub2] [Medline: 23450559]
20. Torous J, Choudhury T, Barnett I, Keshavan M, Kane J. Smartphone relapse prediction in serious mental illness: a pathway towards personalized preventive care. World Psychiatry 2020 Oct;19(3):308-309. [doi: 10.1002/wps.20805]
21. Astill Wright L, Majid M, Shajan G, et al. The user experience of ambulatory assessment and mood monitoring in depression: a systematic review & meta-synthesis. npj Digit Med 2025 Dec 2;8(1):737. [doi: 10.1038/s41746-025-02118-8] [Medline: 41331067]
22. Astill Wright L, Majid M, Moore M, et al. The user experience of ecological momentary assessment and mood monitoring in bipolar disorder: a systematic review and meta-synthesis of qualitative studies. J Med Internet Res 2024;27:e71525-e71525. [doi: 10.2196/71525]
23. Brown S, Ploeger C, Guo B, et al. When a test is more than just a test: findings from patient interviews and survey in the trial of a technology to measure antidepressant medication response (the PReDicT Trial). Compr Psychiatry 2024 Jul;132:152467. [doi: 10.1016/j.comppsych.2024.152467] [Medline: 38608615]
24. Colom F, Lam D. Psychoeducation: improving outcomes in bipolar disorder. Eur Psychiatry 2005 Aug;20(5-6):359-364. [doi: 10.1016/j.eurpsy.2005.06.002] [Medline: 16112848]
25. Van Til K, McInnis MG, Cochran A. A comparative study of engagement in mobile and wearable health monitoring for bipolar disorder. Bipolar Disord 2020 Mar;22(2):182-190. [doi: 10.1111/bdi.12849] [Medline: 31610074]
26. Brown S, Ploeger C, Guo B, et al. When a test is more than just a test: findings from patient interviews and survey in the trial of a technology to measure antidepressant medication response (the PReDicT Trial). Compr Psychiatry 2024 Jul;132:152467. [doi: 10.1016/j.comppsych.2024.152467]
27. Dubad M, Winsper C, Meyer C, Livanou M, Marwaha S. A systematic review of the psychometric properties, usability and clinical impacts of mobile mood-monitoring applications in young people. Psychol Med 2018 Jan;48(2):208-228. [doi: 10.1017/S0033291717001659] [Medline: 28641609]
28. Faurholt-Jepsen M, Munkholm K, Frost M, Bardram JE, Kessing LV. Electronic self-monitoring of mood using IT platforms in adult patients with bipolar disorder: a systematic review of the validity and evidence. BMC Psychiatry 2016 Jan 15;16:7. [doi: 10.1186/s12888-016-0713-0] [Medline: 26769120]
29. Jirotka M, Grimpe B, Stahl B, Eden G, Hartswood M. Responsible research and innovation in the digital age. Commun ACM 2017 Apr 24;60(5):62-68. [doi: 10.1145/3064940]
30. Ortiz A, Maslej MM, Husain MI, Daskalakis ZJ, Mulsant BH. Apps and gaps in bipolar disorder: a systematic review on electronic monitoring for episode prediction. J Affect Disord 2021 Dec;295:1190-1200. [doi: 10.1016/j.jad.2021.08.140]
31. van der Watt ASJ, Odendaal W, Louw K, Seedat S. Distant mood monitoring for depressive and bipolar disorders: a systematic review. BMC Psychiatry 2020 Jul 22;20(1):383. [doi: 10.1186/s12888-020-02782-y] [Medline: 32698802]
32. Astill Wright L, Morriss R. Mood-tracking and self-monitoring in bipolar disorder and depression. NIHR-PROSPERO. 2024. URL: https://www.crd.york.ac.uk/PROSPERO/view/CRD42023396473 [accessed 2025-12-23]
33. Bipolar disorder (update) – review protocols. National Institute for Health and Care Excellence (NICE). 2021. URL: https://www.nice.org.uk/guidance/cg185/documents/bipolar-disorder-update-review-questions2 [accessed 2025-09-17]
34. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
35. Antosik-Wójcińska AZ, Dominiak M, Chojnacka M, et al. Smartphone as a monitoring tool for bipolar disorder: a systematic review including data analysis, machine learning algorithms and predictive modelling. Int J Med Inform 2020 Jun;138:104131. [doi: 10.1016/j.ijmedinf.2020.104131] [Medline: 32305023]
36. Palmier-Claus J, Lobban F, Mansell W, et al. Mood monitoring in bipolar disorder: Is it always helpful? Bipolar Disord 2021 Jun;23(4):429-431. [doi: 10.1111/bdi.13057] [Medline: 33570820]
37. Malhi GS, Hamilton A, Morris G, Mannie Z, Das P, Outhred T. The promise of digital mood tracking technologies: are we heading on the right track? Evid Based Mental Health 2017 Nov;20(4):102-107. [doi: 10.1136/eb-2017-102757]
38. Faurholt-Jepsen M, Geddes JR, Goodwin GM, et al. Reporting guidelines on remotely collected electronic mood data in mood disorder (eMOOD)-recommendations. Transl Psychiatry 2019 Jun 7;9(1):162. [doi: 10.1038/s41398-019-0484-8] [Medline: 31175283]
39. Higgins JPT, Altman DG, Gotzsche PC, et al. The cochrane collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343(oct18 2):d5928-d5928. [doi: 10.1136/bmj.d5928]
40. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. BMJ 2004 Jun 19;328(7454):1490. [doi: 10.1136/bmj.328.7454.1490] [Medline: 15205295]
41. Chandler J, Cumpston M, Li T, Page MJ. In: Welch VA, editor. Cochrane Handbook for Systematic Reviews of Interventions Version 64: Cochrane Collaboration; 2023, Vol. 6:4.
42. Faurholt-Jepsen M, Frost M, Ritz C, et al. Daily electronic self-monitoring in bipolar disorder using smartphones – the MONARCA I trial: a randomized, placebo-controlled, single-blind, parallel group trial. Psychol Med 2015 Oct;45(13):2691-2704. [doi: 10.1017/S0033291715000410]

43. Faurholt-Jepsen M, Frost M, Christensen EM, Bardram JE, Vinberg M, Kessing LV. The effect of smartphone-based monitoring on illness activity in bipolar disorder: the MONARCA II randomized controlled single-blinded trial. Psychol Med 2020 Apr;50(5):838-848. [doi: 10.1017/S0033291719000710] [Medline: 30944054]
44. Faurholt‐Jepsen M, Lindbjerg Tønning M, Fros M, et al. Reducing the rate of psychiatric re‐admissions in bipolar disorder using smartphones—the RADMIS trial. Acta Psychiatr Scand 2021 May;143(5):453-465. [doi: 10.1111/acps.13274]
45. Gliddon E, Cosgrove V, Berk L, et al. A randomized controlled trial of MoodSwings 2.0: an internet-based self-management program for bipolar disorder. Bipolar Disord 2019 Feb;21(1):28-39. [doi: 10.1111/bdi.12669] [Medline: 29931798]
46. Aikens JE, Valenstein M, Plegue MA, et al. Technology-facilitated depression self-management linked with lay supporters and primary care clinics: randomized controlled trial in a low-income sample. Telemed J E Health 2022 Mar;28(3):399-406. [doi: 10.1089/tmj.2021.0042] [Medline: 34086485]
47. Tønning ML, Faurholt-Jepsen M, Frost M, et al. The effect of smartphone-based monitoring and treatment on the rate and duration of psychiatric readmission in patients with unipolar depressive disorder: the RADMIS randomized controlled trial. J Affect Disord 2021 Mar 1;282:354-363. [doi: 10.1016/j.jad.2020.12.141] [Medline: 33421863]
48. Hunkeler EM, Hargreaves WA, Fireman B, et al. A web-delivered care management and patient self-management program for recurrent depression: a randomized trial. Psychiatr Serv 2012 Nov;63(11):1063-1071. [doi: 10.1176/appi.ps.005332011] [Medline: 22983558]
49. Bilderbeck AC, Atkinson LZ, McMahon HC, et al. Psychoeducation and online mood tracking for patients with bipolar disorder: A randomised controlled trial. J Affect Disord 2016 Nov;205:245-251. [doi: 10.1016/j.jad.2016.06.064]
50. Denicoff KD, Ali SO, Sollinger AB, Smith-Jackson EE, Leverich GS, Post RM. Utility of the daily prospective National Institute of Mental Health Life-Chart Method (NIMH-LCM-p) ratings in clinical trials of bipolar disorder. Depress Anxiety 2002;15(1):1-9. [doi: 10.1002/da.1078] [Medline: 11816046]
51. Faurholt-Jepsen M, Frost M, Ritz C, et al. Daily electronic self-monitoring in bipolar disorder using smartphones - the MONARCA I trial: a randomized, placebo-controlled, single-blind, parallel group trial. Psychol Med 2015 Oct;45(13):2691-2704. [doi: 10.1017/S0033291715000410] [Medline: 26220802]
52. Lauder S, Chester A, Castle D, et al. A randomized head to head trial of MoodSwings.net.au: an Internet based self-help program for bipolar disorder. J Affect Disord 2015 Jan 15;171:13-21. [doi: 10.1016/j.jad.2014.08.008] [Medline: 25282145]
53. Castle D, White C, Chamberlain J, et al. Group-based psychosocial intervention for bipolar disorder: randomised controlled trial. Br J Psychiatry 2010 May;196(5):383-388. [doi: 10.1192/bjp.bp.108.058263] [Medline: 20435965]
54. Petzold J, Mayer-Pelinski R, Pilhatsch M, et al. Short group psychoeducation followed by daily electronic self-monitoring in the long-term treatment of bipolar disorders: a multicenter, rater-blind, randomized controlled trial. Int J Bipolar Disord 2019 Nov 4;7(1):23. [doi: 10.1186/s40345-019-0158-8] [Medline: 31680193]
55. van den Berg KC, Hendrickson AT, Hales SA, Voncken M, Keijsers GPJ. Comparing the effectiveness of imagery focussed cognitive therapy to group psychoeducation for patients with bipolar disorder: a randomised trial. J Affect Disord 2023 Jan;320:691-700. [doi: 10.1016/j.jad.2022.09.160]
56. Goldberg JF, Bowden CL, Calabrese JR, et al. Six-month prospective life charting of mood symptoms with lamotrigine monotherapy versus placebo in rapid cycling bipolar disorder. Biol Psychiatry Cogn Neurosci Neuroimaging 2008 Jan;63(1):125-130. [doi: 10.1016/j.biopsych.2006.12.031]
57. Langosch JM, Drieling T, Biedermann NC, et al. Efficacy of quetiapine monotherapy in rapid-cycling bipolar disorder in comparison with sodium valproate. J Clin Psychopharmacol 2008;28(5):555-560. [doi: 10.1097/JCP.0b013e318185e75f]
58. Lieberman DZ, Kelly TF, Douglas L, Goodwin FK. A randomized comparison of online and paper mood charts for people with bipolar disorder. J Affect Disord 2010 Jul;124(1-2):85-89. [doi: 10.1016/j.jad.2009.10.019]
59. Depp CA, Kim DH, de Dios LV, Wang V, Ceglowski J. A pilot study of mood ratings captured by mobile phone versus paper-and-pencil mood charts in bipolar disorder. J Dual Diagn 2012 Jan 1;8(4):326-332. [doi: 10.1080/15504263.2012.723318] [Medline: 23646035]
60. Leverich GS, Altshuler LL, Frye MA, et al. Risk of switch in mood polarity to hypomania or mania in patients with bipolar depression during acute and continuation trials of venlafaxine, sertraline, and bupropion as adjuncts to mood stabilizers. Am J Psychiatry 2006 Feb;163(2):232-239. [doi: 10.1176/appi.ajp.163.2.232] [Medline: 16449476]
61. Pahwa M, McElroy SL, Priesmeyer R, et al. KIOS: a smartphone app for self-monitoring for patients with bipolar disorder. Bipolar Disord 2024 Feb;26(1):84-92. [doi: 10.1111/bdi.13362] [Medline: 37340215]
62. Faurholt-Jepsen M, Brage S, Vinberg M, et al. Electronic monitoring of psychomotor activity as a supplementary objective measure of depression severity. Nord J Psychiatry 2015 Feb;69(2):118-125. [doi: 10.3109/08039488.2014.936501] [Medline: 25131795]
63. Faurholt-Jepsen M, Frost M, Martiny K, et al. Reducing the rate and duration of Re-ADMISsions among patients with unipolar disorder and bipolar disorder using smartphone-based monitoring and treatment – the RADMIS trials: study protocol for two randomized controlled trials. Trials 2017 Dec;18(1). [doi: 10.1186/s13063-017-2015-3]
64. Tønning ML, Faurholt-Jepsen M, Frost M, et al. The effect of smartphone-based monitoring and treatment on the rate and duration of psychiatric readmission in patients with unipolar depressive disorder: the RADMIS randomized controlled trial. J Affect Disord 2021 Mar 1;282(Supplement 1):354-363. [doi: 10.1016/j.jad.2020.12.141] [Medline: 33421863]

65. Jonathan G, Goulding E, Dopke C, et al. Understand the mechanisms of behavior change in livewell: a smartphone intervention for bipolar disorder. Presented at: 22nd Annual ISBD Conference 20/20: Vision for Bipolar Disorder and Depression; Jun 18-21, 2020. [doi: 10.1111/bdi.12939]
66. Saunders KEA, Bilderbeck AC, Panchal P, Atkinson LZ, Geddes JR, Goodwin GM. Experiences of remote mood and activity monitoring in bipolar disorder: a qualitative study. Eur psychiatr 2017;41(1):115-121. [doi: 10.1016/j.eurpsy.2016.11.005]
67. Wellness tracker. Depression and Bipolar Support Alliance. URL: https://www.dbsalliance.org/wellness/wellness-toolbox/wellness-tracker/ [accessed 2025-08-17]
68. Daily mood chart. Black Dog Institute. URL: https://www.blackdoginstitute.org.au/wp-content/uploads/2020/04/19-dailymoodchart.pdf [accessed 2025-08-17]
69. Successfully manage bipolar disorder, anxiety, and depression. MoodTracker. URL: https://www.moodtracker.com/ [accessed 2025-08-17]
70. Linardon J, Torous J, Firth J, Cuijpers P, Messer M, Fuller - Tyszkiewicz M. Current evidence on the efficacy of mental health smartphone apps for symptoms of depression and anxiety. a meta - analysis of 176 randomized controlled trials. World Psychiatry 2024 Feb;23(1):139-149. [doi: 10.1002/wps.21183]
71. Astill Wright L, Bakstein E, Saunders K, Guo B, Morriss R. Performance of active and passive ambulatory assessment measures and mood monitoring in bipolar disorder: a systematic review. Int J Bipolar Disord (forthcoming). Preprint posted online on 2024. [doi: 10.1186/s40345-025-00407-5]
72. Astill Wright L, Roe J, Guo B, Morriss R. Attrition, adherence, and compliance in mood monitoring and ambulatory assessment studies for depression and bipolar disorder: systematic review and meta-analysis. JMIR Ment Health (forthcoming) 2026. [doi: 10.2196/83765]
73. Koh J, Tng GYQ, Hartanto A. Potential and pitfalls of mobile mental health apps in traditional treatment: an umbrella review. J Pers Med 2022 Aug 25;12(9):1376. [doi: 10.3390/jpm12091376] [Medline: 36143161]
74. Wang M, Liu Q, Yang X, et al. Relationship of insight to neurocognitive function and risk of recurrence in depression: a naturalistic follow-up study. Front Psychiatry 2023;14:1084993. [doi: 10.3389/fpsyt.2023.1084993]
75. aan het Rot M, Hogenelst K, Schoevers RA. Mood disorders in everyday life: a systematic review of experience sampling and ecological momentary assessment studies. Clin Psychol Rev 2012 Aug;32(6):510-523. [doi: 10.1016/j.cpr.2012.05.007]
76. Western MJ, Smit ES, Gültzow T, et al. Bridging the digital health divide: a narrative review of the causes, implications, and solutions for digital health inequalities. Health Psychol Behav Med 2025;13(1):2493139. [doi: 10.1080/21642850.2025.2493139] [Medline: 40276490]
77. Gnanapragasam SN, Astill Wright L, Pemberton M, Bhugra D. Outside/inside: social determinants of mental health. Ir J Psychol Med 2023 Mar;40(1):63-73. [doi: 10.1017/ipm.2021.49]
78. Salem M, Robenson J. The impact of socioeconomic factors on mental health: a conceptual framework. Cureus 2025;17(7):e88244. [doi: 10.7759/cureus.88244]
79. Astill Wright L, Majid M, Moore M, et al. The user experience of ambulatory assessment and mood monitoring in bipolar disorder: systematic review and meta-synthesis of qualitative studies. J Med Internet Res 2025 Oct 17;27:e71525. [doi: 10.2196/71525] [Medline: 41105870]

Abbreviations

BD: bipolar disorder
EMA: ecological momentary assessment
IVR: interactive voice response
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
PROSPERO: International Prospective Register of Systematic Reviews
RCT: randomized controlled trial
RCT: randomized controlled trial
SMD: standardized mean difference

Edited by J Torous; submitted 12.Sep.2025; peer-reviewed by B Najand, E Goulding, P Harvey; revised version received 25.Nov.2025; accepted 26.Nov.2025; published 07.Jan.2026.

Please cite as:
Astill Wright L, Shajan G, Purewal D, Stone J, Majid M, Guo B, Morriss R
Mood Monitoring, Mood Tracking, and Ambulatory Assessment Interventions in Depression and Bipolar Disorder: Systematic Review and Meta-Analysis of Randomized Controlled Trials
JMIR Ment Health 2026;13:e84020
URL: https://mental.jmir.org/2026/1/e84020
doi:10.2196/84020

Dropout, Attrition, Adherence, and Compliance in Mood Monitoring and Ambulatory Assessment Studies for Depression and Bipolar Disorder: Systematic Review and Meta-Analysis

Laurence Astill Wright[1,2], MRCPsych; James Roe[1], PhD; Boliang Guo[1], PhD; Richard Morriss[1], PhD

[1]Institute of Mental Health, University of Nottingham, Jubilee Campus, University of Nottingham Innovation Park, Triumph Road, Nottingham, United Kingdom

[2]Centre for Academic Mental Health, Population Health Sciences, University of Bristol, Bristol, United Kingdom

Corresponding Author:
Laurence Astill Wright, MRCPsych
Institute of Mental Health, University of Nottingham, Jubilee Campus, University of Nottingham Innovation Park, Triumph Road, Nottingham, United Kingdom

Abstract

Background: Ambulatory assessment and mood monitoring are different methods that can use novel technology to deliver a more efficient, flexible, and usable method of clinical outcome assessment compared with established measures of behavior and mood. Concerns have been raised around attrition in and adherence to these new protocols, particularly over the medium to long term by people with mood disorders.

Objective: This systematic review and meta-analysis assessed attrition from and adherence to active and passive ambulatory assessment and mood monitoring protocols by people with bipolar disorder and depression over the medium and long term.

Methods: Randomized controlled trials and nonrandomized studies were identified and rated for risk of bias. Adherence and attrition data were pooled to calculate effect sizes. We analyzed specific factors that we hypothesized a priori could affect the prevalences of attrition and adherence by means of subgroup meta-analysis or metaregression modeling.

Results: We evaluated 77 mood tracking or ambulatory assessment studies including 17,123 participants. Pooled adherence was 0.64% (95% CI 0.59%-0.70%; *P*<.001), and pooled attrition was 0.28% (95% CI 0.22%-0.34%; *P*<.001). Three factors had a statistically significant subgroup difference for adherence: The presence of financial incentives increased adherence, and the presence of mood monitoring reminders and a higher study risk of bias decreased adherence. Four factors had a statistically significant subgroup difference for attrition: Digital mood monitoring decreased attrition versus analogue studies, but mood monitoring reminders, mood monitoring versus other protocols, and a high risk of study bias increased attrition. These analyses, however, were vulnerable to confounding by study design and protocol design. Attrition rates were not reported by 17 studies (17/77, 22%), and 20 studies (20/77, 26%) did not report adherence rates. Most studies had a low-to-moderate risk of bias, but heterogeneity was very high. Only 16 studies reported adherence systematically.

Conclusions: Reporting of attrition and adherence to ambulatory assessments was not systematic nor universal, and until it is, analyses are unlikely to demonstrate clear conclusions. We found very high heterogeneity and evidence of publication bias, and this limited the certainty of our conclusions. Financial incentives may increase adherence, and attrition may be lower in digital than analogue studies of mood monitoring. There was no statistically significant difference in adherence and attrition between studies of passive and active ambulatory assessments. Reminders of mood monitoring increased attrition and decreased adherence, but the results may be confounded by longer length of follow-up versus other studies.

Trial Registration: PROSPERO CRD42023396473; https://www.crd.york.ac.uk/PROSPERO/view/CRD42023396473

(JMIR Ment Health 2026;13:e83765) doi:10.2196/83765

KEYWORDS

bipolar; depression; EMA; ecological momentary assessment; ambulatory assessment; mood tracking; mood monitoring; self-monitoring; adherence; attrition; dropout; drop out; compliance

Introduction

Technological developments, such as wearables and smartphone sensors, have enhanced the capacity for mood monitoring through both active (manual input) and passive (automated) data collection [1]. Although these tools offer promising applications in mood disorders (particularly for addressing the limitations of traditional measurement tools for capturing rapidly shifting mood change [2]), smartphone-based ambulatory assessments may pose specific risks and challenges, especially for vulnerable populations [3]. The definition of ambulatory assessment encompasses a wide range of methods that leverage mobile technologies to repeatedly study individuals—often in real time and in their natural environments [4]. This includes mood monitoring (repeatedly tracking one's mood over time), remote measurement technologies (defined as wearable devices that passively collect data without any active user input), and ecological momentary assessment (EMA), which is defined as a type of assessment involving more intensive, frequent self-reports throughout the day in real time and "in the moment" [5].

This review explored dropout, attrition, and adherence associated with mood monitoring, mood tracking, and ambulatory assessment by individuals with unipolar depression and bipolar disorder. This study focused on studies exploring mood monitoring (defined as regularly tracking and appraising one's mood), and many of the studies included here also fall under the similar categories of EMA, remote measurement technologies, or ambulatory assessment depending on the frequency, manner, and technology used to track mood. There is also considerable overlap between these definitions (defined in the previous paragraph). Here, mood monitoring was used both as an intervention (in randomized and nonrandomized studies) and as a method for measuring outcomes (again in randomized and nonrandomized studies).

Because of this, this study focused specifically on individuals with unipolar depression and bipolar disorder. This is due to the evidence suggesting that this group faces distinct risks and challenges that may cause issues with attrition and adherence [6,7]. Mood disorders represent promising conditions for mood monitoring, as they are characterized by subtle mood shifts over time that may respond well to early intervention. They also represent a large global source of disability with depression as the leading cause [8]. Both people with depression and people with bipolar disorder (particularly those who experience psychosis) may have particularly poor adherence rates; therefore, improving this may be vital to realizing some of the benefits that ambulatory assessment approaches could offer for understanding and treating these conditions [9,10].

Due to the perceived burden of some ambulatory assessment protocols [6], concerns have been raised about their usability and acceptability, particularly over longer periods of time in clinical populations [9]. Therefore, this study focused on 3 months or longer. There are likely to be factors that could be adjusted to optimize usability and acceptability and thus adherence and attrition. This could improve the overall quality of the data collected and thus improve the research or interventions that derive from the data. Optimizing adherence and attrition is also of benefit to clinical research and could considerably decrease sample sizes required for future ambulatory assessment research [11].

Ambulatory assessment protocols are highly heterogenous and, even within the same discipline, use a wide variety of different methodologies [5]. These vary in terms of the type and frequency of assessment schedule as well as the technology used to collect the data and whether the assessment is also used as an intervention. Because of this heterogeneity, previous reviews have attempted to untangle which factors affect adherence and attrition, but these have only been in short-term use rather than medium or long-term assessment. This medium or long-term use may be necessary to evaluate the fluctuations in mood that occur over time, particularly in depression and bipolar disorder [9,11]. Longer-term data are likely to be fundamental for digital phenotyping and offering personalized psychiatry, which could, for example, identify early warning signs and prevent or predict relapse [12], aims that are particularly relevant for mood disorder research [13,14].

There are several reasons why adherence to long-term ambulatory assessment use may be different to that of shorter protocols. Ambulatory assessment protocol adherence tends to decrease over the duration of the assessment and by people with higher levels of negative affect (eg, mood disorders), but again, this has not been assessed over longer time frames [9,15]. Previous work has identified that studies can improve adherence by offering financial incentives, minimizing the quantity of assessments, and the assessments occurring at a prespecified time [9,11]. Other studies have not identified any significant predictors of attrition [16]. It remains unclear if these factors also apply to medium or long-term approaches.

This review aimed to assess attrition and adherence over the medium and long term to ambulatory assessment and mood monitoring protocols and consider the factors that may affect these (eg, diagnosis, financial incentives, remote or in-person enrollment, reminders to complete the mood assessment, frequency of mood assessment).

Methods

We used a Cochrane Handbook for Systematic Reviews of Interventions–based methodology. We completed a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) checklist. The study was pre-registered in accordance with the International Prospective Register of Systematic Reviews (PROSPERO: CRD42023396473 [17]).

Inclusion Criteria

Included studies met the following criteria: included self-monitoring, mood monitoring, or repeated symptom assessment with people with bipolar disorder or depression using an interventional or observational design over a minimum period of 3 months, with rating of symptoms weekly at a minimum. Operationally, for this review, we defined this repeated symptom assessment as either mood monitoring or ambulatory assessment (as defined in the previous section—with all mood monitoring conducted as an ambulatory assessment but not all ambulatory assessment being mood monitoring). We chose this broad definition because many ambulatory assessment protocols have the potential to be used for mood monitoring as an intervention or for research outcomes, even if this was not the intended aim of the study, although all included studies used the repeated symptom assessment either as an intervention or research outcome.

The studies should have either used a validated measure of mood or validated the chosen measure with a validated mood measure. The studies could be published in any language and could be digital or nondigital, although we acknowledged that the majority of studies would utilize digital technologies. We included nonrandomized or randomized studies with 20 or more participants with bipolar disorder or depression [18]. We searched the gray literature (eg, conference abstracts, dissertations, policy literature, reports via ProQuest and Google Scholar—full details are in the next section) for unpublished studies that were eligible for inclusion.

Search Strategy and Selection Criteria

The complete search strategy is listed in Multimedia Appendix 1. We searched Ovid MEDLINE, EMBASE, PsychINFO, SCOPUS, IEE Xplore, Proquest SciTech Collection, Proquest dissertations and theses global, and Google Scholar using the search terms. The initial search was conducted on March 3, 2023, and updated on October 28, 2024. All abstracts were appraised by 2 independent screeners, any disagreements were discussed, and a consensus was reached, with adjudication by a third independent screener if required. We acquired the full text of any potentially relevant papers, and if we were unable to source the full text of the study, we then contacted the corresponding author to request the paper. To determine if potentially relevant studies met the inclusion criteria, the full text was reviewed separately by 2 authors, again with discussion and consensus with a third reviewer if necessary. All papers for inclusion were reference checked along with relevant systematic reviews [19-28]. Key authors were also emailed to see if any ongoing unpublished studies could be included.

Data Extraction

Data were extracted by 2 independent reviewers from studies meeting the inclusion criteria using identical data extraction forms. Irregularities in the data extraction were discussed, and any discrepancy was resolved with discussion.

In this review, we defined attrition as the number of participants who had withdrawn from the study and no longer contributed data after randomization or baseline assessment. We defined adherence as the number of ambulatory assessments completed by those who remained in the study. When calculating attrition and adherence, we included anyone randomized into the trial, even if they did not subsequently complete any or many mood assessments. This was because preliminary analyses showed the reporting of the exact number of people who took part in the mood monitoring was poor, and our aim was to create pragmatic attrition and adherence values that would be useful to the power calculations of future ambulatory assessment studies. Using the number randomized was likely to affect the calculated estimates and was likely to increase the attrition rate by increasing the denominator size. For example, this may have included individuals without any actual use of the monitoring tool but who were randomized to the intervention arm. Even if not reflective of the adherence and attrition of highly motivated individuals, nonetheless, this is to give a pragmatic figure for the design of future studies. This estimate then gave a pragmatic figure for the design of future studies that use an optimal intention-to-treat design rather than estimates based on selection of more motivated and adherent participants.

Attrition and adherence were calculated using what we judged to be the most systematic and pragmatic metric available (eg, the longest follow-up available using systematically acquired data). Not all studies reported raw figures, and some only reported attrition %, which we included when available. For only a minority of studies, adherence was reported systematically using device-generated data. Many studies did not report this, and some studies reported the number of participants completing the study as the adherence: We included these less accurate metrics if there was not a more systematic metric reported. For many of these metrics, we were required to email the authors to request the attrition and adherence data due to no reporting in the manuscript. For attrition, using the number randomized as the denominator may have underestimated the attrition, as many of these individuals would have never used the ambulatory assessment protocol. On the other hand, including some nonsystematically gathered data likely distorts the pooled estimate (possibly by underestimating attrition and adherence but this is not certain), but we did, however, control for this in a sensitivity analysis documented in the following sections.

Assessment of Study Bias

For observational studies, the Cochrane Collaboration tool for assessing the risk of bias in nonrandomized studies of interventions [29] was used. For randomized controlled trials (RCTs), we used the Cochrane risk of bias tool for randomized trials for each study [30]. Risk of bias was assessed by 2 independent reviewers, and any disagreement was resolved via discussion.

Synthesis of Results

We grouped studies together, where possible, according to the variable assessed (eg, adherence or attrition) and pooled the data in a meta-analysis. We conducted subgroup analysis on specific factors that we hypothesized a priori could affect the prevalences of attrition and adherence. Results of each primary study were pooled using the inverse variance weighted approach with a random or fixed effects model, informed by examining the between-studies heterogeneities. Stata metan code was used to perform the analysis for proportion data, and metareg code was used for metaregression modeling.

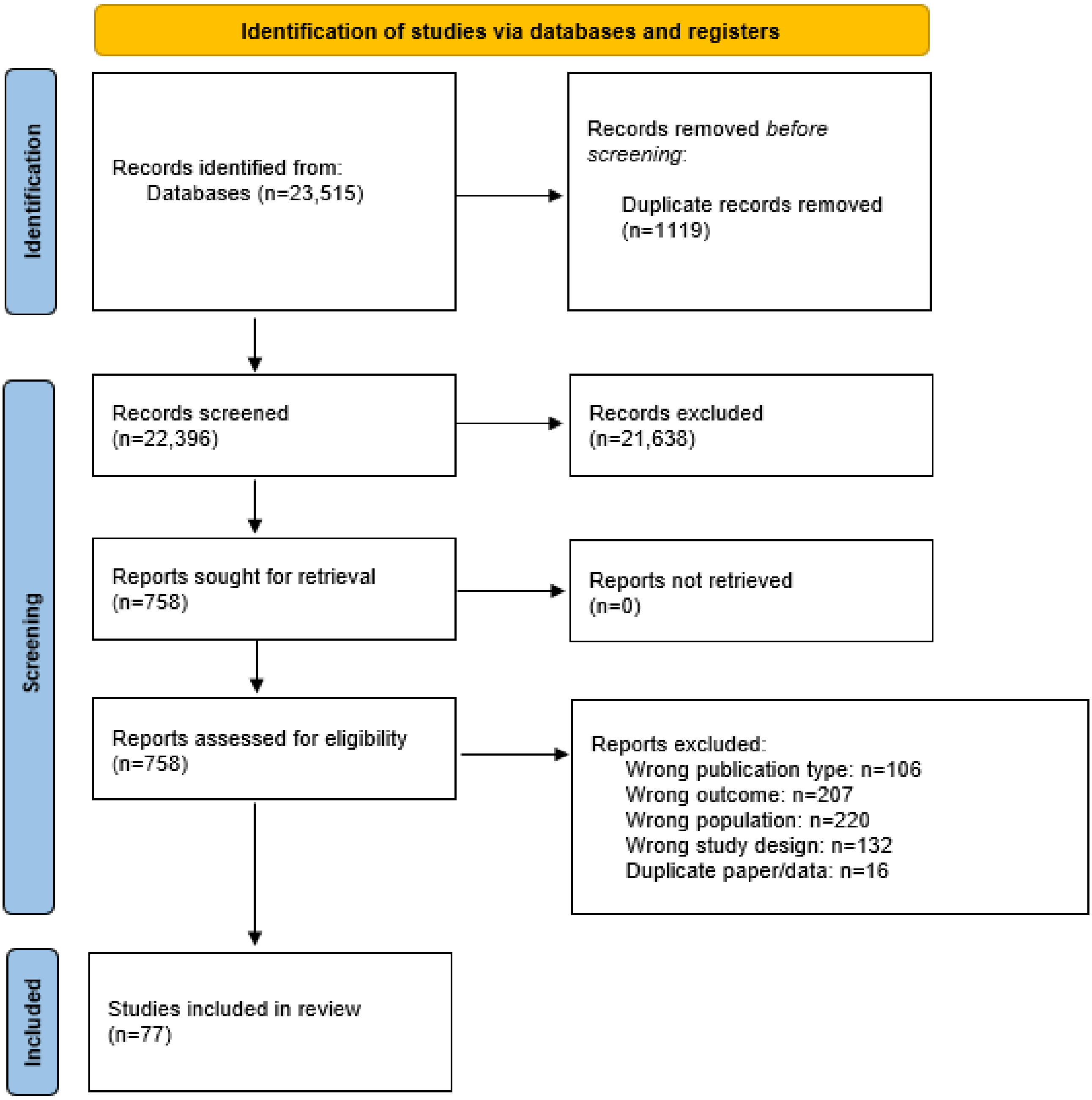
Results

Search Results

The search identified 23,515 papers (Figure 1). No studies not published in English met the inclusion criteria. Following title and abstract screening, 21,638 articles were excluded, resulting in a total of 758 papers being reviewed in full. A total of 77 papers met the eligibility criteria and were included in the review. The 77 studies included 17,123 participants: 17 studies did not report attrition rates, 20 studies did not report adherence rates, 34 studies did not report if financial incentives were offered, and 25 studies did not report the recruitment method. The average mood tracking or ambulatory assessment duration was 6.9 months. Reporting of adherence was not universal nor systematic: Of the 57 studies that reported adherence, 34 studies did not specify how the adherence statistic was calculated, and 23 systematically used device-generated data to calculate this.

Tables S1 and S2 in Multimedia Appendix 2 display the detailed characteristics of the studies and the mood monitoring protocols used. These tables are stratified by clinical condition and type of study (eg, randomized or nonrandomized).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flow diagram of included studies.



We included mixed samples if they included 20 or more participants with either depression or bipolar disorder. In the depression studies, this included 899 participants with primary diagnoses other than depression or healthy controls of a total of 9310 participants included. In bipolar disorder studies, there were 83 participants with primary diagnoses other than bipolar disorder or healthy controls of a total of 7813 included. In total, 5.7% (977/17,123) of the total participants did not have a primary diagnosis of depression or bipolar disorder.

We included relapse prevention or maintenance trials, as this is a core element of treatment of bipolar disorder and depression, which are usually relapsing and remitting illnesses. With studies of variable follow-up durations, we included the attrition or adherence from the latest available outcome, and this was a pragmatic decision around the design of future studies that will require long-term follow-ups due to the lifelong nature of many mood disorders.

Risk of Bias Assessments

The methodological quality of these studies was variable, but most were considered of low-to-moderate risk of bias (Table S3 in Multimedia Appendix 2).

Meta-Analysis

As planned a priori, we conducted a subgroup analysis of 5 different factors for both adherence and attrition reported in Table S1 in Multimedia Appendix 2: type of mood monitoring, mood disorder, RCT versus nonrandomized study, digitization of mood monitoring, presence of reminders to complete the mood monitoring, in-person or remote enrollment, and the presence of financial incentives.

Meta-analytic pooling of adherence (0.64%, 95% CI 0.59%-0.70%; P<.001) and attrition (0.28%, 95% CI 0.22%-0.34%; P<.001) is reported in Figures 2 and 3, respectively. For adherence, heterogeneity was high (Cochran Q_{90}=10,052.49; P<.001; I^2=99.1%). Likewise for attrition, heterogeneity was high (Cochran Q_{73}=6615.17; P<.001; I^2=98.9%). Because of the pronounced heterogeneity, we used a random effects model to pool the data. Both of the funnel plots for the total prevalences of adherence and attrition demonstrated visual asymmetry and thus suggested possible publication bias or small study effects in the meta-analytic estimate (Figures S1 and S3 in Multimedia Appendix 2). The Egger test demonstrated that there were no small-study effects (Figures S2 and S4 in Multimedia Appendix 2) for both the adherence and attrition analyses.

Figure 2. Forest plot of the prevalence of adherence with percentage weighting [31-96].

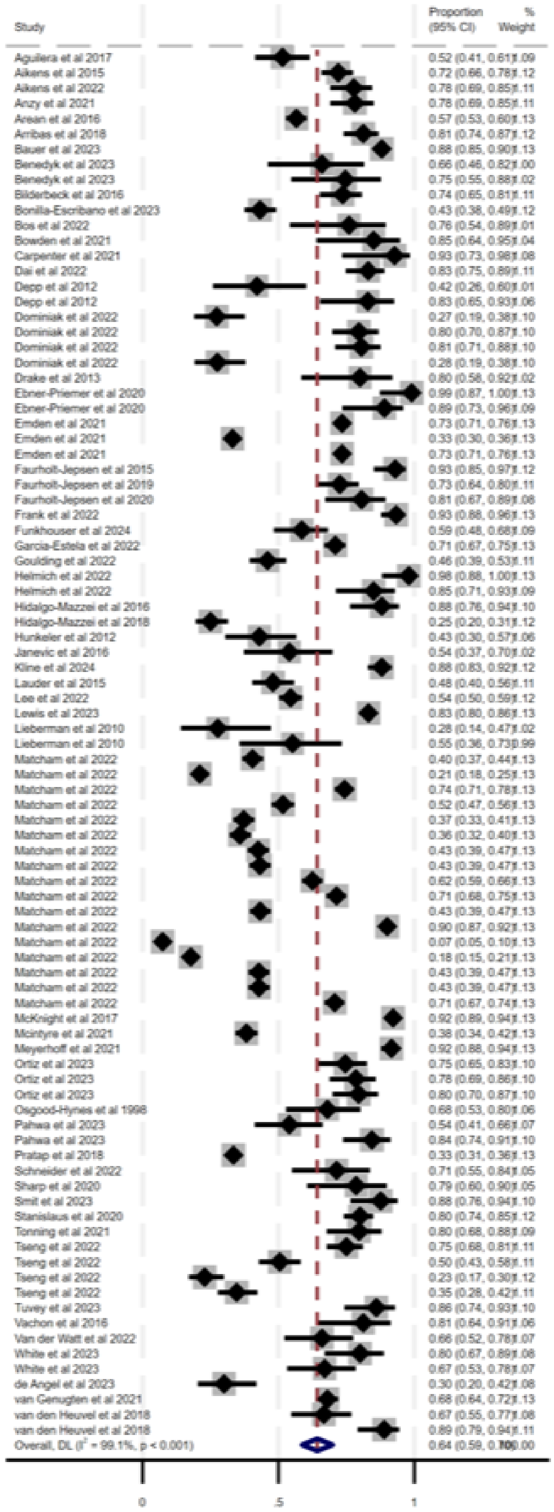
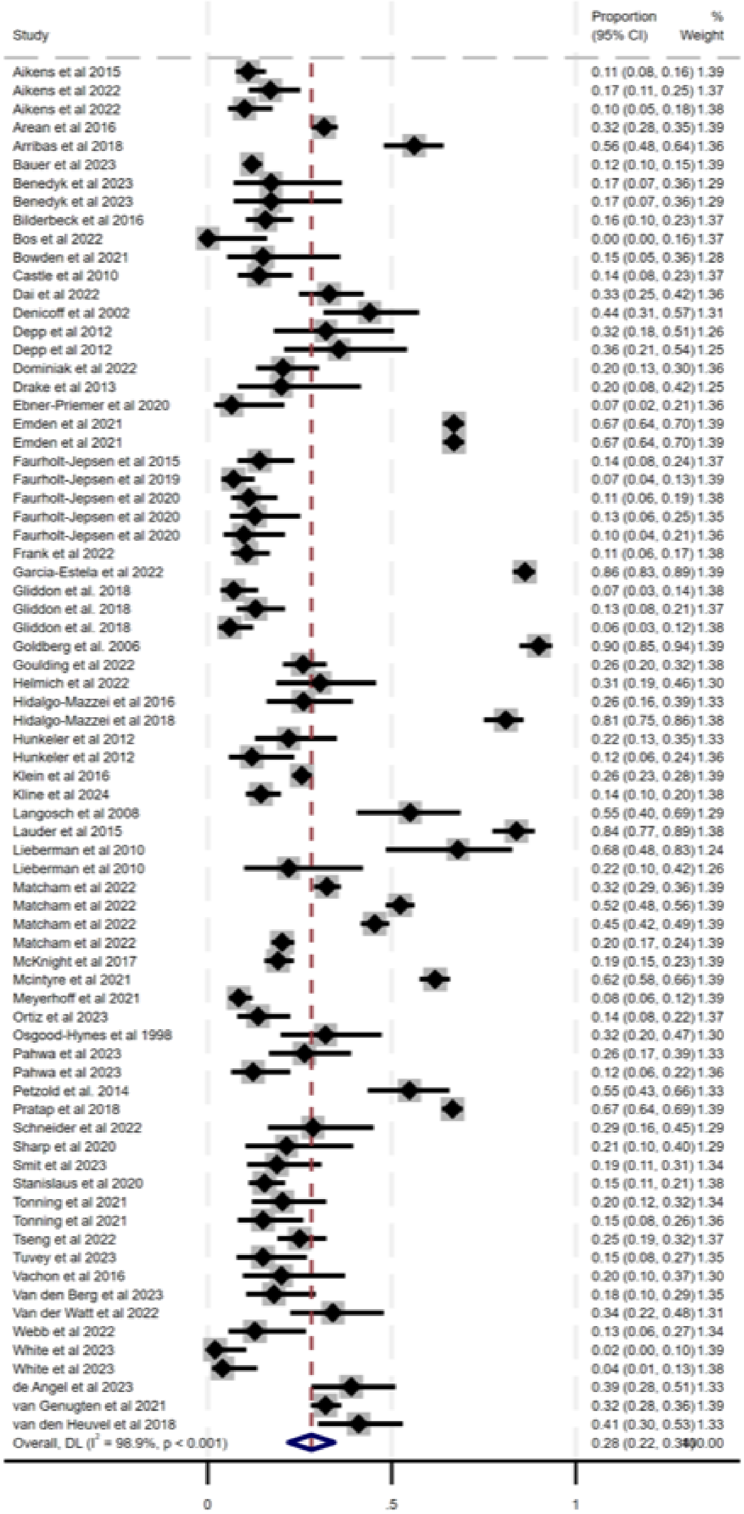


Figure 3. Forest plot of the prevalence of attrition with percentage weighting [31-41,43,44,46-52,54-57,59-63,64,65,66,67,68,70,71-73,74-79,80,81-86,87,88,89-91,92-96].



We report subgroup analyses in Figures S5-S16 in Multimedia Appendix 2. The interpretation of any differences here should not be interpreted causally due to the limitations of the data; likewise, any notable absence of effect should not be interpreted as conclusive equivalence due to potential power limitations. Only 3 factors had a statistically significant subgroup difference for adherence: the presence of financial incentives, mood monitoring reminders, and study risk of bias. Studies without financial incentives had lower adherence (0.65%, 95% CI 0.52%-0.79%) than those offering financial incentives (0.77%, 95% CI 0.66%-0.88%; between-group Q_2=6.38; P=.04). There was not a statistically significant effect of financial incentives on attrition, however: studies without financial incentives: 0.22% (95% CI 0.07%-0.37%); studies with financial incentives: 0.21% (95% CI 0.10%-0.31%; between-group Q_2=3.91; P=.14). Studies with mood monitoring reminders had lower adherence (0.61%, 95% CI 0.54%-0.67%) than studies without mood monitoring reminders (0.78%, 95% CI 0.69%-0.88%; between-group Q_2=11.25; *P*=.004).

A statistically significant subgroup difference was also not present for adherence for studies that systematically reported adherence via device-level data compared with those that did not: studies that did not systematically report adherence: 0.67% (95% CI 0.59%-0.75%); studies that systematically reported adherence: 0.62% (95% CI 0.54%-0.69%; between-group Q_1=1; *P*=.32).

A statistically significant subgroup difference for attrition was present for 5 factors: presence of mood monitoring reminders, digital vs analogue mood monitoring, mood monitoring study or non-mood monitoring control, and study risk of bias. Studies with mood monitoring reminders had higher attrition (0.30%, 95% CI 0.22%-0.37%) than studies without mood monitoring reminders (0.27%, 95% CI 0.16%-0.38%; between-group Q_2=56.80; *P*<.001). Analogue studies had higher attrition (0.40%, 95% CI 0.16%-0.64%) than digital studies (0.28%, 95% CI 0.21%-0.35%; between-group Q_2=29.66; *P*<.001). Mood monitoring studies had higher attrition (0.29%, 95% CI 0.23%-0.36%) than non-mood monitoring control groups (0.09%, 95% CI 0.06%-0.12%; between-group Q_2=31.94; *P*<.001).

There is likely an element of confounding by study design and protocol burden in these subgroup analyses, as studies with EMA reminders (n=23) reported a longer mean follow-up duration (mean 11.02, SD 8.5 months) than studies without EMA reminders (n=86; mean 7.48, SD 4.22 months).

There was no statistically significant difference in adherence nor attrition between active and passive mood monitoring (Figures S5 and S12 in Multimedia Appendix 2), in-person versus remote enrollment or recruitment (Figures S11 and S18 in Multimedia Appendix 2), and depression versus bipolar disorder (Figures S6 and S13 in Multimedia Appendix 2).

There were statistically significant subgroup differences in both adherence and attrition between studies deemed to have a high risk of bias (≤4 areas of low risk of bias as classified on the Cochrane risk of bias tool) versus those deemed to have a low risk of bias (5 or 6 areas of low risk of bias). High-risk studies had lower adherence (0.60%, 95% CI 0.53%-0.68%) than low-risk studies (0.70%, 95% CI 0.64%-0.76%; between-group Q_1=3.93; *P*=.047). High-risk studies had higher attrition (0.38%, 95% CI 0.29%-0.46%) than low-risk studies (0.19%, 95% CI 0.14%-0.24%; between-group Q_1=13.75; *P*<.001).

Metaregression

As the heterogeneity was very high, we conducted a metaregression to assess potential sources of this heterogeneity. We explored the effects of financial incentives, ambulatory assessment reminders, digital or analogue protocol, active or passive ambulatory assessment, gender, age, and ambulatory assessment duration on heterogeneity. The residual heterogeneity remained very high despite this, and the full results are reported in Table S4 in Multimedia Appendix 2.

Notably, active ambulatory assessments had a statistically significant higher attrition than nonambulatory assessment–based controls (odds ratio [OR] 1.23, 95% CI 1.00-1.51; *P*=.046). However, attrition rates for studies using both active and passive ambulatory assessments were not statistically significant compared with the control group (OR 1.14, 95% CI 0.92-1.42; *P*=.23). Likewise, attrition rates for passive ambulatory assessment studies were not statistically significantly different than the control groups (OR 1.23, 95% CI 0.97-1.57; *P*=.09). This suggests studies using just active ambulatory assessment measures may have a uniquely high burden for participants.

There was a statistically significant negative effect of ambulatory assessment duration on adherence rates (OR 0.99, 95% CI 0.99-1.00; *P*=.01), and adherence declined as assessment duration increased.

There was a statistically significant effect of the number of low-risk domains in a study leading to lower attrition (OR 0.94, 95% CI 0.90-0.98; *P*=.002) but not adherence (OR 1.02, 95% CI 0.98-1.07; *P*=.26).

Discussion

Summary of Findings

This meta-analysis examined long-term adherence and attrition in mood monitoring or ambulatory assessment studies by people with depression and bipolar disorder and, to our best knowledge, is the first review to do so. We documented larger improvements in attrition in digital studies than in analogue studies, in adherence in studies with financial incentives versus nonfinancial incentives, and increased attrition or decreased adherence in studies with mood monitoring reminders versus those without reminders. There was also improved adherence or attrition in higher-quality studies. Crucially, there were no differences in adherence and attrition between studies of passive and active mood monitoring, although active ambulatory assessments had higher attrition than non-mood monitoring controls. However, we noted there was a very high level of heterogeneity in the meta-analysis, with some evidence of publication bias, which may mean that studies showing no difference in attrition or adherence were not necessarily reported, and no clear picture has emerged from our results. The high heterogeneity was not due to the reasons explored in the metaregression: age, gender, assessment duration, active or passive assessment, digital or analogue, or assessment reminders or financial incentives. Reporting of adherence and attrition was not universal nor systematic, and until it is, analyses are unlikely to provide any clear conclusions.

Context of Findings

The mood monitoring adherence statistics we report here (many of which were EMA studies) are lower than those in studies of dropout and adherence with EMA protocols in both clinical and nonclinical populations [11,97-99]. They are also lower than the recommended rate of 80% for EMA studies [100]. Missing data, particularly higher than suggested recommended rates, can limit the conclusions that are drawn from statistical inference and have a significant effect on the study's statistical power

[98]. Furthermore, research suggests that missing data in EMA studies are not at random and thus more likely to introduce nonresponse bias—for example, participants are more likely to miss certain assessments at specific times (eg, when delivering childcare at a specific time in the evening) [101]. The rates of adherence reported here are thus concerning to the methodological validity of these mood monitoring studies, some of which are also EMA studies. This is further demonstrated by the higher attrition in mood monitoring groups when compared with non-mood monitoring controls, raising the question of whether mood monitoring is itself aversive.

The follow-up length that we assessed was also much longer than other studies and other reviews, and this may partially explain the lower adherence. Here, the average mood monitoring or ambulatory assessment protocol length was approximately 7 months, while many reviews have only assessed adherence over 7 days. Long-term studies generally demonstrate higher attrition than short-term studies [102]. Furthermore, clinical populations may have higher attrition rates and lower adherence than nonclinical populations. This is particularly true of people with mood disorders who can struggle with drive, motivation, and organization. The attrition rate was comparable to other studies of bipolar disorder [103] or depression and clinical trials of smartphone apps [104]. This was reassuring considering the increased participant burden reported in qualitative studies and suggests that mood monitoring studies, with regard to attrition and adherence, may not be particularly different or uniquely stressful than studies using established measures of assessment.

Protocol Design

Crucially, we did not demonstrate better adherence and attrition in passive mood monitoring studies than with active protocols. This was in contrast to qualitative research suggesting passive mood monitoring is preferred, as it is perceived as being less intrusive (and has less participation burden) by people with bipolar disorder and depression [6]. This may be due to people not adhering to or dropping out from mood monitoring studies for reasons other than the frequent completion of self-report measures. Some of these concerns driving nonadherence and dropout may be around data security, not feeling that they themselves are receiving a benefit, or the presence of adverse effects, as opposed to the type of mood monitoring or the way this is delivered [6,105].

Financial incentives did not appear to affect attrition but were associated with improved adherence. This was in keeping with previous research with similar findings [11], finding 12% higher adherence if financial incentives were given (77%) compared with when financial incentives were not given (65%). This is a very large improvement in adherence, particularly over many months. There are, however, concerns that paying individuals to participate in research may exert undue influence and coercion [106] and potentially select individuals who are more motivated by the financial reward than the potential for digital technology to improve their mental disorder via mood monitoring or to contribute to research via ambulatory assessment—both of which likely require motivation and commitment [6]. This may thus create selection biases. This also may be especially problematic in studies where reward responsivity is being measured since financial incentives might be perceived as a reward and therefore a confounder. Participants should also be given the choice to not accept a financial reward if they choose; some people find such rewards coercive or forming an obligation when they wish to take part from free will.

Mood monitoring reminders were associated with increased attrition and decreased adherence. It is possible that a higher number of notifications and reminders increases participant burden and leads to dropout by some individuals or that those studies with more burdensome protocols also include reminders. There is likely an element of confounding by study design and protocol burden in these analyses. Furthermore, this is in contrast to other findings, where increased participant burden (eg, with active mood monitoring) was not associated with an increased dropout versus passive mood monitoring (which is often designed to minimize this burden). One possible explanation for the lack of difference in dropout despite an increased participant burden is that active mood monitoring directly asks the user their opinion and makes them aware of their mood versus passive monitoring, which sits in the background and provides no immediate feedback to the participant. This active feedback-driven process may compensate for the increased burden.

Reporting Standards

Reporting of attrition (60/77, 78% of studies) and adherence (57/77, 74% of studies) was low, suggesting that studies should report these metrics routinely, particularly considering the importance of these metrics to mood monitoring data validity. The number of studies reporting these metrics in the published manuscripts was even lower; we contacted a large number of authors individually to find much of this information, which was often difficult to locate in the first instance. Only a small minority of studies reported adherence in a systematic way from device data—most did not clarify the nature of the data they reported (eg, self-report adherence, objective app data).

Researchers have previously produced important guidelines on the reporting of mood monitoring data (eg, eMOOD guidelines [26]). Although these do report the importance of good adherence data, they do not mention the systematic underreporting of adherence and attrition statistics and provide no factual basis for the statement that adherence to mood monitoring is often low. The quantified evidence summarized here provides a basis for specific recommendations around the reporting of adherence and attrition data, which has previously been sparse. We suggest that preference be given to systematically collected device-level data that allow direct comparison between protocols. Acceptable attrition and adherence thresholds should be prespecified based on power calculations and the intended purpose of the ambulatory assessment. Power calculations for sample size should consider what effects acceptable attrition and adherence might have on outcome; power calculations without these considerations might then be considered incomplete and probably underpowered.

The reporting of attrition should be in accordance with CONSORT (Consolidated Standards of Reporting Trials) for randomized studies [107] and alternatives for nonrandomized studies [108]. This will allow researchers to assess attrition based on the number of participants enrolled or those with actual use of the ambulatory assessment, as these figures are likely to tell us different things and inform future studies in different ways. Researchers should report attrition based on the number of participants enrolled as an intention-to-treat analysis, as this will generate a more generalizable result for clinical practice. They should also report a per-protocol analysis of those who attempted to use the assessment as intended, as this will give an indication of attrition by those motivated and willing to undertake these assessments at the beginning of the study. These align with current guidelines for interventions in general in RCTs and nonrandomized studies. We make further broader recommendations incorporating a broad range of qualitative and quantitative evidence that puts poor adherence and attrition reporting in the wider ambulatory assessment or mood monitoring context [109].

Lower risk-of-bias studies had considerably less attrition and higher adherence. Higher-quality studies may be more rigorously conducted, and this may lead to improved follow-up rates and more participant-friendly protocols (eg, via more qualitative development of the protocol or lived experience input) to which it is easier to adhere. Therefore, the overall risk of bias of a study may be a marker of a higher-quality study that is less onerous for participants. These attrition and adherence results, even for high-quality studies, suggest that there is a high level of attrition and low level of adherence to data collection that are hard to address with methods that rely on longer-term active data collection alone. Further investigation on the optimal mix of methods of digital collection (eg, active and passive with optimal methods of support and reward) is merited .

Participant Engagement

Future mood monitoring studies with mood disorders should prioritize consistent reporting of adherence and attrition and use financial incentives to support participant engagement. This study failed to find a multitude of specific factors that improve attrition and adherence, aside from financial incentives and reminders (which surprisingly had a negative effect on adherence and attrition). This thus raises questions about additional methods of improving the acceptability and usability of these protocols. Many of these factors are likely to be protocol specific and will likely require lived experience input, qualitative research, and ethical implementation to create a protocol that has high acceptability and usability, particularly with long-term studies. Some mood monitoring protocols reported attrition as high as 92%, and others reported adherence as low as 18%. It remains unclear what is a realistic attrition or adherence rate for mood monitoring studies with people with mood disorders, and this may rarely reach the recommended 80%, particularly for individuals with a higher severity of illness.

Strengths and Limitations

The strengths of this review are that we evaluated studies using mood monitoring over a minimum period of 3 months. We argue that 3 months of follow-up was required to accurately assess mood, attrition, and adherence in a clinically meaningful way to assist the design of medium or long-term trials in the future. The average duration of follow-up was much longer than other reviews—approximately 7 months—a more clinically material length of time. We did not include many mood monitoring procedures that utilized shorter follow-up and are further away from implementation as interventions or as research tools. This allows us to make clear conclusions about mood monitoring methods that have the ability to affect the current delivery of mental health trials. One limitation of this review is that the vast majority of trials had issues with risk of bias, specifically around blinding of the intervention or mood monitoring procedure, but this is an issue with trials of psychological treatments more broadly where perfect blinding is not possible. We were also not able to include all studies in the meta-analysis due to underreporting of key data, which is in itself both a limitation and an important finding highlighting the lack of standardization in the reporting of key mood monitoring metrics. Furthermore, there was large statistical heterogeneity when aggregating data, and this reflects the heterogeneity of the wide-ranging mood monitoring protocols (eg, type of mood disorder, population, duration of mood monitoring, and intensity of mood monitoring); despite using a random-effects model, this may still limit the generalizability of our findings, mask important contextual differences, and limit the comparability of the summary estimates across studies. In addition, our analyses should not be interpreted causally due to these limitations, and they were vulnerable to confounding by study design and protocol burden.

Conclusions

To conclude, this meta-analysis examined long-term adherence and attrition in mood monitoring studies by people with mood disorders. Attrition and adherence were lower for people with bipolar disorder and depression than for other nonclinical or clinical populations but comparable to other non-mood monitoring studies with people with mood disorders. We demonstrated that some key factors may improve adherence and attrition (crucially, there were no differences in adherence and attrition between studies of passive and active mood monitoring), but the certainty of our findings is limited due to the lack of systematic and universal reporting of adherence and attrition in ambulatory assessment studies. These findings should inform the design of future mood monitoring protocols—prioritizing systematic and universal reporting of adherence and attrition—and be interpreted alongside qualitative findings to optimize real-world acceptability and utility. Until reporting of adherence and attrition improves, further analyses are unlikely to provide any clear conclusions.

Acknowledgments

We would like to thank the University of Nottingham Library Service for their assistance with formulating the search terms.

Funding

This work was supported by a Wellcome Trust Fellowship to LAW. RM has received other National Institute of Health and Care Research (NIHR) funding for research on interventions for depression and bipolar disorder, including the lead for the Nottingham NIHR Biomedical Research Centre mental health and technology theme, the NIHR Applied Research Collaborations (ARC) East Midlands Mental Health and Well-being Theme, and the NIHR MindTech Health Technology Collaborative.

Authors' Contributions

Conceptualization: LAW, RM
Formal analysis: LAW, JR, BG
Investigation: LAW, JR
Writing – original draft: LAW
Writing – review & editing: LAW, RM, JM, BG

Conflicts of Interest

RM has received funding from Novartis to serve on a data management and ethics committee for two trials on the treatment of depression.

Multimedia Appendix 1
Search strategy
[DOCX File, 31 KB - mental_v13i1e83765_app1.docx]

Multimedia Appendix 2
Study characteristics and risk of bias assessments.
[DOCX File, 58321 KB - mental_v13i1e83765_app2.docx]

Checklist 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) checklist.
[PDF File, 83 KB - mental_v13i1e83765_app3.pdf]

References

1. Matcham F, Barattieri di San Pietro C, Bulgari V, et al. Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol. BMC Psychiatry 2019 Feb 18;19(1):72. [doi: 10.1186/s12888-019-2049-z] [Medline: 30777041]
2. Gromatsky M, Sullivan SR, Spears AP, et al. Ecological momentary assessment (EMA) of mental health outcomes in veterans and servicemembers: a scoping review. Psychiatry Res 2020 Oct;292:113359. [doi: 10.1016/j.psychres.2020.113359] [Medline: 32777594]
3. 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 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
4. Trull TJ, Ebner-Priemer U. Ambulatory assessment. Annu Rev Clin Psychol 2013;9:151-176. [doi: 10.1146/annurev-clinpsy-050212-185510] [Medline: 23157450]
5. aan het Rot M, Hogenelst K, Schoevers RA. Mood disorders in everyday life: a systematic review of experience sampling and ecological momentary assessment studies. Clin Psychol Rev 2012 Aug;32(6):510-523. [doi: 10.1016/j.cpr.2012.05.007] [Medline: 22721999]
6. Astill Wright L, Majid M, Moore M, et al. The user experience of ambulatory assessment and mood monitoring in bipolar disorder: systematic review and meta-synthesis of qualitative studies. J Med Internet Res 2025 Oct 17;27:e71525. [doi: 10.2196/71525] [Medline: 41105870]
7. van Genugten CR, Schuurmans J, Lamers F, et al. Experienced burden of and adherence to smartphone-based ecological momentary assessment in persons with affective disorders. J Clin Med 2020 Jan 23;9(2):322. [doi: 10.3390/jcm9020322] [Medline: 31979340]
8. Friedrich MJ. Depression is the leading cause of disability around the world. JAMA 2017 Apr 18;317(15):1517. [doi: 10.1001/jama.2017.3826]
9. Vachon H, Viechtbauer W, Rintala A, Myin-Germeys I. Compliance and retention with the experience sampling method over the continuum of severe mental disorders: meta-analysis and recommendations. J Med Internet Res 2019 Dec 6;21(12):e14475. [doi: 10.2196/14475] [Medline: 31808748]
10. Sokolovsky AW, Mermelstein RJ, Hedeker D. Factors predicting compliance to ecological momentary assessment among adolescent smokers. Nicotine Tob Res 2014 Mar;16(3):351-358. [doi: 10.1093/ntr/ntt154] [Medline: 24097816]

11. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. Assessment 2023 Apr;30(3):825-846. [doi: 10.1177/10731911211067538] [Medline: 35016567]
12. Baumeister H, Montag C. Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics: Springer Nature; 2019:291. [doi: 10.1007/978-3-030-31620-4]
13. Morriss RK, Faizal MA, Jones AP, Williamson PR, Bolton C, McCarthy JP. Interventions for helping people recognise early signs of recurrence in bipolar disorder. Cochrane Database Syst Rev 2007 Jan 24;2007(1):CD004854. [doi: 10.1002/14651858.CD004854.pub2] [Medline: 17253526]
14. Morriss RK, Bolton CA, McCarthy JP, Marshall M, Williamson PR, Jones AP. Interventions for helping people recognise early signs of recurrence in depression. Cochrane Database Syst Rev 2018(10). [doi: 10.1002/14651858.CD004855.pub2]
15. Fuller-Tyszkiewicz M, Skouteris H, Richardson B, Blore J, Holmes M, Mills J. Does the burden of the experience sampling method undermine data quality in state body image research? Body Image 2013 Sep;10(4):607-613. [doi: 10.1016/j.bodyim.2013.06.003] [Medline: 23856302]
16. Bell IH, Eisner E, Allan S, et al. Methodological characteristics and feasibility of ecological momentary assessment studies in psychosis: a systematic review and meta-analysis. Schizophr Bull 2024 Mar 7;50(2):238-265. [doi: 10.1093/schbul/sbad127] [Medline: 37606276]
17. Astill Wright L, Morriss R. Mood-tracking and self-monitoring in bipolar disorder and depression. PROSPERO. 2023. URL: https://www.crd.york.ac.uk/PROSPERO/view/CRD42023396473 [accessed 2025-12-24]
18. Bipolar Disorder (Update) – Review Protocols. National Institute for Health and Care Excellence. 2021. URL: https://www.nice.org.uk/guidance/cg185/documents/bipolar-disorder-update-review-questions2 [accessed 2025-12-24]
19. Antosik-Wójcińska AZ, Dominiak M, Chojnacka M, et al. Smartphone as a monitoring tool for bipolar disorder: a systematic review including data analysis, machine learning algorithms and predictive modelling. Int J Med Inform 2020 Jun;138:104131. [doi: 10.1016/j.ijmedinf.2020.104131] [Medline: 32305023]
20. Dubad M, Winsper C, Meyer C, Livanou M, Marwaha S. A systematic review of the psychometric properties, usability and clinical impacts of mobile mood-monitoring applications in young people. Psychol Med 2018 Jan;48(2):208-228. [doi: 10.1017/S0033291717001659] [Medline: 28641609]
21. Palmier-Claus J, Lobban F, Mansell W, et al. Mood monitoring in bipolar disorder: Is it always helpful? Bipolar Disord 2021 Jun;23(4):429-431. [doi: 10.1111/bdi.13057] [Medline: 33570820]
22. Faurholt-Jepsen M, Munkholm K, Frost M, Bardram JE, Kessing LV. Electronic self-monitoring of mood using IT platforms in adult patients with bipolar disorder: a systematic review of the validity and evidence. BMC Psychiatry 2016 Jan 15;16:7. [doi: 10.1186/s12888-016-0713-0] [Medline: 26769120]
23. van der Watt ASJ, Odendaal W, Louw K, Seedat S. Distant mood monitoring for depressive and bipolar disorders: a systematic review. BMC Psychiatry 2020 Jul 22;20(1):383. [doi: 10.1186/s12888-020-02782-y] [Medline: 32698802]
24. Ortiz A, Maslej MM, Husain MI, Daskalakis ZJ, Mulsant BH. Apps and gaps in bipolar disorder: a systematic review on electronic monitoring for episode prediction. J Affect Disord 2021 Dec 1;295:1190-1200. [doi: 10.1016/j.jad.2021.08.140] [Medline: 34706433]
25. Malhi GS, Hamilton A, Morris G, Mannie Z, Das P, Outhred T. The promise of digital mood tracking technologies: are we heading on the right track? Evid Based Mental Health 2017 Nov;20(4):102-107. [doi: 10.1136/eb-2017-102757]
26. Faurholt-Jepsen M, Geddes JR, Goodwin GM, et al. Reporting guidelines on remotely collected electronic mood data in mood disorder (eMOOD)-recommendations. Transl Psychiatry 2019 Jun 7;9(1):162. [doi: 10.1038/s41398-019-0484-8] [Medline: 31175283]
27. Wenze SJ, Miller IW. Use of ecological momentary assessment in mood disorders research. Clin Psychol Rev 2010 Aug;30(6):794-804. [doi: 10.1016/j.cpr.2010.06.007] [Medline: 20619520]
28. Koenders MA, Nolen WA, Giltay EJ, Hoencamp E, Spijker AT. The use of the prospective NIMH Life Chart Method as a bipolar mood assessment method in research: a systematic review of different methods, outcome measures and interpretations. J Affect Disord 2015 Apr 1;175:260-268. [doi: 10.1016/j.jad.2015.01.005] [Medline: 25658502]
29. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
30. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343(oct18 2):d5928. [doi: 10.1136/bmj.d5928] [Medline: 22008217]
31. Hidalgo-Mazzei D, Mateu A, Reinares M, et al. Psychoeducation in bipolar disorder with a SIMPLe smartphone application: feasibility, acceptability and satisfaction. J Affect Disord 2016 Aug;200:58-66. [doi: 10.1016/j.jad.2016.04.042] [Medline: 27128358]
32. van den Heuvel S, Meije D, Regeer EJ, Sinnema H, Riemersma RF, Kupka RW. The user experiences and clinical outcomes of an online personal health record to support self-management of bipolar disorder: a pretest-posttest pilot study. J Affect Disord 2018 Oct;238:261-268. [doi: 10.1016/j.jad.2018.05.069]
33. Van der Watt ASJ, Dalvie N, Seedat S. Weekly telephone mood monitoring is associated with decreased suicidality and improved sleep quality in a clinical sample. Psychiatry Res 2022 Nov;317:114821. [doi: 10.1016/j.psychres.2022.114821] [Medline: 36088835]

34. Drake G, Csipke E, Wykes T. Assessing your mood online: acceptability and use of Moodscope. Psychol Med 2013 Jul;43(7):1455-1464. [doi: 10.1017/S0033291712002280] [Medline: 23149120]
35. Hidalgo-Mazzei D, Reinares M, Mateu A, et al. OpenSIMPLe: a real-world implementation feasibility study of a smartphone-based psychoeducation programme for bipolar disorder. J Affect Disord 2018 Dec 1;241:436-445. [doi: 10.1016/j.jad.2018.08.048] [Medline: 30145515]
36. García-Estela A, Cantillo J, Angarita-Osorio N, et al. Real-world implementation of a smartphone-based psychoeducation program for bipolar disorder: observational ecological study. J Med Internet Res 2022 Feb 2;24(2):e31565. [doi: 10.2196/31565] [Medline: 35107440]
37. de Angel V, Adeleye F, Zhang Y, et al. The feasibility of implementing remote measurement technologies in psychological treatment for depression: mixed methods study on engagement. JMIR Ment Health 2023 Jan 24;10:e42866. [doi: 10.2196/42866] [Medline: 36692937]
38. Osgood-Hynes DJ, Greist JH, Marks IM, et al. Self-administered psychotherapy for depression using a telephone-accessed computer system plus booklets. J Clin Psychiatry 1998 Jul 15;59(7):358-365. [doi: 10.4088/JCP.v59n0704]
39. Sharp KJ, South CC, Chin Fatt C, Trivedi MH, Rethorst CD. Pilot studies to evaluate feasibility of a physical activity intervention for persons with depression. J Sport Exerc Psychol ;42(6):443-451. [doi: 10.1123/jsep.2019-0248]
40. Pahwa M, McElroy SL, Priesmeyer R, et al. KIOS: a smartphone app for self-monitoring for patients with bipolar disorder. Bipolar Disord 2024 Feb;26(1):84-92 [FREE Full text] [doi: 10.1111/bdi.13362] [Medline: 37340215]
41. Bowden CL, Priesmeyer R, Tohen M, et al. Development of a patient-centered software system to facilitate effective management of bipolar disorder. Psychopharmacol Bull 2021 Mar 16;51(2):8-19. [doi: 10.64719/pb.4404] [Medline: 34092819]
42. Janevic MR, Aruquipa Yujra AC, Marinec N, et al. Feasibility of an interactive voice response system for monitoring depressive symptoms in a lower-middle income Latin American country. Int J Ment Health Syst 2016;10(1):59. [doi: 10.1186/s13033-016-0093-3] [Medline: 27688798]
43. Matcham F, Leightley D, Siddi S, et al. Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study. BMC Psychiatry 2022 Feb 21;22(1):136. [doi: 10.1186/s12888-022-03753-1] [Medline: 35189842]
44. Faurholt-Jepsen M, Frost M, Ritz C, et al. Daily electronic self-monitoring in bipolar disorder using smartphones - the MONARCA I trial: a randomized, placebo-controlled, single-blind, parallel group trial. Psychol Med 2015 Oct;45(13):2691-2704. [doi: 10.1017/S0033291715000410] [Medline: 26220802]
45. Bonilla-Escribano P, Ramírez D, Baca-García E, Courtet P, Artés-Rodríguez A, López-Castromán J. Multidimensional variability in ecological assessments predicts two clusters of suicidal patients. Sci Rep 2023 Mar 2;13(1):3546. [doi: 10.1038/s41598-023-30085-1] [Medline: 36864070]
46. Meyerhoff J, Liu T, Kording KP, et al. Evaluation of changes in depression, anxiety, and social anxiety using smartphone sensor features: longitudinal cohort study. J Med Internet Res 2021 Sep 3;23(9):e22844. [doi: 10.2196/22844] [Medline: 34477562]
47. Faurholt-Jepsen M, Frost M, Christensen EM, Bardram JE, Vinberg M, Kessing LV. The effect of smartphone-based monitoring on illness activity in bipolar disorder: the MONARCA II randomized controlled single-blinded trial. Psychol Med 2020 Apr;50(5):838-848. [doi: 10.1017/S0033291719000710] [Medline: 30944054]
48. Lauder S, Chester A, Castle D, et al. A randomized head to head trial of MoodSwings.net.au: an internet based self-help program for bipolar disorder. J Affect Disord 2015 Jan 15;171:13-21. [doi: 10.1016/j.jad.2014.08.008] [Medline: 25282145]
49. Goulding EH, Dopke CA, Rossom R, Jonathan G, Mohr D, Kwasny MJ. Effects of a smartphone-based self-management intervention for individuals With bipolar disorder on relapse, symptom burden, and quality of life: a randomized clinical trial. JAMA Psychiatry 2023 Feb 1;80(2):109-118. [doi: 10.1001/jamapsychiatry.2022.4304] [Medline: 36542401]
50. Bilderbeck AC, Atkinson LZ, McMahon HC, et al. Psychoeducation and online mood tracking for patients with bipolar disorder: a randomised controlled trial. J Affect Disord 2016 Nov 15;205:245-251. [doi: 10.1016/j.jad.2016.06.064] [Medline: 27454410]
51. White KM, Carr E, Leightley D, et al. Engagement with a remote symptom-tracking platform among participants with major depressive disorder: randomized controlled trial. JMIR Mhealth Uhealth 2024 Jan 19;12:e44214. [doi: 10.2196/44214] [Medline: 38241070]
52. Tønning ML, Faurholt-Jepsen M, Frost M, et al. The effect of smartphone-based monitoring and treatment on the rate and duration of psychiatric readmission in patients with unipolar depressive disorder: the RADMIS randomized controlled trial. J Affect Disord 2021 Mar 1;282:354-363. [doi: 10.1016/j.jad.2020.12.141] [Medline: 33421863]
53. Aguilera A, Bruehlman-Senecal E, Demasi O, Avila P. Automated text messaging as an adjunct to cognitive behavioral therapy for depression: a clinical trial. J Med Internet Res 2017 May 8;19(5):e148. [doi: 10.2196/jmir.6914] [Medline: 28483742]
54. Hunkeler EM, Hargreaves WA, Fireman B, et al. A web-delivered care management and patient self-management program for recurrent depression: a randomized trial. Psychiatr Serv 2012 Nov;63(11):1063-1071. [doi: 10.1176/appi.ps.005332011] [Medline: 22983558]

55. Aikens JE, Trivedi R, Heapy A, Pfeiffer PN, Piette JD. Potential impact of incorporating a patient-selected support person into mHealth for depression. J Gen Intern Med 2015 Jun;30(6):797-803. [doi: 10.1007/s11606-015-3208-7] [Medline: 25666218]
56. Helmich MA, Smit AC, Bringmann LF, et al. Detecting impending symptom transitions using early-warning signals in individuals receiving treatment for depression. Clin Psychol Sci 2023 Nov;11(6):994-1010. [doi: 10.1177/21677026221137006]
57. Dominiak M, Kaczmarek-Majer K, Antosik-Wójcińska AZ, et al. Behavioral and self-reported data collected from smartphones for the assessment of depressive and manic symptoms in patients with bipolar disorder: prospective observational study. J Med Internet Res 2022 Jan 19;24(1):e28647. [doi: 10.2196/28647] [Medline: 34874015]
58. Anýž J, Bakštein E, Dally A, et al. Validity of the Aktibipo self-rating questionnaire for the digital self-assessment of mood and relapse detection in patients with bipolar disorder: instrument validation study. JMIR Ment Health 2021 Aug 9;8(8):e26348. [doi: 10.2196/26348] [Medline: 34383689]
59. Schneider J, Bakštein E, Kolenič M, et al. Motor activity patterns can distinguish between interepisode bipolar disorder patients and healthy controls. CNS Spectr 2022 Feb;27(1):82-92. [doi: 10.1017/S1092852920001777] [Medline: 32883376]
60. Benedyk A, Moldavski A, Reichert M, et al. Initial response to the COVID-19 pandemic on real-life well-being, social contact and roaming behavior in patients with schizophrenia, major depression and healthy controls: a longitudinal ecological momentary assessment study. Eur Neuropsychopharmacol 2023 Apr;69:79-83. [doi: 10.1016/j.euroneuro.2023.01.008] [Medline: 36791492]
61. Emden D, Goltermann J, Dannlowski U, Hahn T, Opel N. Technical feasibility and adherence of the Remote Monitoring Application in Psychiatry (ReMAP) for the assessment of affective symptoms. J Affect Disord 2021 Nov 1;294:652-660. [doi: 10.1016/j.jad.2021.07.030] [Medline: 34333173]
62. Lieberman DZ, Kelly TF, Douglas L, Goodwin FK. A randomized comparison of online and paper mood charts for people with bipolar disorder. J Affect Disord 2010 Jul;124(1-2):85-89. [doi: 10.1016/j.jad.2009.10.019] [Medline: 19896202]
63. Smit AC, Snippe E, Bringmann LF, Hoenders HJR, Wichers M. Transitions in depression: if, how, and when depressive symptoms return during and after discontinuing antidepressants. Qual Life Res 2023 May;32(5):1295-1306. [doi: 10.1007/s11136-022-03301-0] [Medline: 36418524]
64. Aikens JE, Valenstein M, Plegue MA, et al. Technology-facilitated depression self-management linked with lay supporters and primary care clinics: randomized controlled trial in a low-income sample. Telemed J E Health 2022 Mar;28(3):399-406. [doi: 10.1089/tmj.2021.0042] [Medline: 34086485]
65. Arean PA, Hallgren KA, Jordan JT, et al. The use and effectiveness of mobile apps for depression: results from a fully remote clinical trial. J Med Internet Res 2016 Dec 20;18(12):e330. [doi: 10.2196/jmir.6482] [Medline: 27998876]
66. Perez Arribas I, Goodwin GM, Geddes JR, Lyons T, Saunders KEA. A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. Transl Psychiatry 2018 Dec 13;8(1):274. [doi: 10.1038/s41398-018-0334-0] [Medline: 30546013]
67. Bauer M, Glenn T, Alda M, et al. Longitudinal digital mood charting in bipolar disorder: experiences with ChronoRecord over 20 years. Pharmacopsychiatry 2023 Sep;56(5):182-187. [doi: 10.1055/a-2156-5667] [Medline: 37678394]
68. Bos FM, Schreuder MJ, George SV, et al. Anticipating manic and depressive transitions in patients with bipolar disorder using early warning signals. Int J Bipolar Disord 2022 Apr 9;10(1):12. [doi: 10.1186/s40345-022-00258-4] [Medline: 35397076]
69. Carpenter L, Hindley L, Gonsalves M, Schatten H, Brown J, Tirrell E. App-based ecological momentary assessment and symptom-adaptive scheduling of TMS maintenance treatments to prevent depressive episode recurrence. Brain Stimul 2021 Nov;14(6):1705. [doi: 10.1016/j.brs.2021.10.373]
70. Castle D, White C, Chamberlain J, et al. Group-based psychosocial intervention for bipolar disorder: randomised controlled trial. Br J Psychiatry 2010 May;196(5):383-388. [doi: 10.1192/bjp.bp.108.058263] [Medline: 20435965]
71. Dai R. Smart sensing and clinical predictions with wearables: from physiological signals to mental health. McKelvey School of Engineering Theses & Dissertations. 2022. URL: https://openscholarship.wustl.edu/eng_etds/779/ [accessed 2026-01-10]
72. Denicoff KD, Ali SO, Sollinger AB, Smith-Jackson EE, Leverich GS, Post RM. Utility of the daily prospective National Institute of Mental Health Life-Chart Method (NIMH-LCM-p) ratings in clinical trials of bipolar disorder. Depress Anxiety 2002;15(1):1-9 [FREE Full text] [doi: 10.1002/da.1078] [Medline: 11816046]
73. Depp CA, Kim DH, de Dios LV, Wang V, Ceglowski J. A pilot study of mood ratings captured by mobile phone versus paper-and-pencil mood charts in bipolar disorder. J Dual Diagn 2012 Jan 1;8(4):326-332. [doi: 10.1080/15504263.2012.723318] [Medline: 23646035]
74. Ebner-Priemer UW, Mühlbauer E, Neubauer AB, et al. Digital phenotyping: towards replicable findings with comprehensive assessments and integrative models in bipolar disorders. Int J Bipolar Disord 2020 Nov 17;8(1):35. [doi: 10.1186/s40345-020-00210-4] [Medline: 33211262]
75. Faurholt-Jepsen M, Lindbjerg Tønning M, Fros M, et al. Reducing the rate of psychiatric re-admissions in bipolar disorder using smartphones-the RADMIS trial. Acta Psychiatr Scand 2021 May;143(5):453-465 [FREE Full text] [doi: 10.1111/acps.13274] [Medline: 33354769]

76. Frank E, Wallace ML, Matthews MJ, et al. Personalized digital intervention for depression based on social rhythm principles adds significantly to outpatient treatment. Front Digit Health 2022;4:870522. [doi: 10.3389/fdgth.2022.870522] [Medline: 36120713]
77. Funkhouser CJ, Trivedi E, Li LY, et al. Detecting adolescent depression through passive monitoring of linguistic markers in smartphone communication. J Child Psychol Psychiatry 2024 Jul;65(7):932-941 [FREE Full text] [doi: 10.1111/jcpp.13931] [Medline: 38098445]
78. Gliddon E, Cosgrove V, Berk L, et al. A randomized controlled trial of MoodSwings 2.0: an internet-based self-management program for bipolar disorder. Bipolar Disord 2019 Feb;21(1):28-39 [FREE Full text] [doi: 10.1111/bdi.12669] [Medline: 29931798]
79. Goldberg JF, Bowden CL, Calabrese JR, et al. Six-month prospective life charting of mood symptoms with lamotrigine monotherapy versus placebo in rapid cycling bipolar disorder. Biol Psychiatry 2008 Jan 1;63(1):125-130. [doi: 10.1016/j.biopsych.2006.12.031] [Medline: 17543894]
80. Klein JP, Berger T, Schröder J, et al. Effects of a psychological internet intervention in the treatment of mild to moderate depressive symptoms: results of the EVIDENT study, a randomized controlled trial. Psychother Psychosom 2016;85(4):218-228. [doi: 10.1159/000445355] [Medline: 27230863]
81. Kline EA, Lekkas D, Bryan A, et al. The role of borderline personality disorder traits in predicting longitudinal variability of major depressive symptoms among a sample of depressed adults. J Affect Disord 2024 Oct 15;363:492-500. [doi: 10.1016/j.jad.2024.07.104] [Medline: 39029689]
82. Langosch JM, Drieling T, Biedermann NC, et al. Efficacy of quetiapine monotherapy in rapid-cycling bipolar disorder in comparison with sodium valproate. J Clin Psychopharmacol 2008 Oct;28(5):555-560. [doi: 10.1097/JCP.0b013e318185e75f] [Medline: 18794653]
83. Lee HJ, Cho CH, Lee T, et al. Prediction of impending mood episode recurrence using real-time digital phenotypes in major depression and bipolar disorders in South Korea: a prospective nationwide cohort study. Psychol Med 2023 Sep;53(12):5636-5644. [doi: 10.1017/S0033291722002847] [Medline: 36146953]
84. Lewis KJS, Tilling K, Gordon-Smith K, et al. The dynamic interplay between sleep and mood: an intensive longitudinal study of individuals with bipolar disorder. Psychol Med 2023 Jun;53(8):3345-3354. [doi: 10.1017/S0033291721005377] [Medline: 35074035]
85. McKnight RF, Bilderbeck AC, Miklowitz DJ, Hinds C, Goodwin GM, Geddes JR. Longitudinal mood monitoring in bipolar disorder: course of illness as revealed through a short messaging service. J Affect Disord 2017 Dec 1;223:139-145. [doi: 10.1016/j.jad.2017.07.029] [Medline: 28753472]
86. McIntyre RS, Lee Y, Rong C, et al. Ecological momentary assessment of depressive symptoms using the mind.me application: convergence with the Patient Health Questionnaire-9 (PHQ-9). J Psychiatr Res 2021 Mar;135:311-317. [doi: 10.1016/j.jpsychires.2021.01.012] [Medline: 33540296]
87. Ortiz A, Park Y, Gonzalez-Torres C, et al. Predictors of adherence to electronic self-monitoring in patients with bipolar disorder: a contactless study using growth mixture models. Int J Bipolar Disord 2023 May 17;11(1):18. [doi: 10.1186/s40345-023-00297-5] [Medline: 37195477]
88. Petzold J, Mayer-Pelinski R, Pilhatsch M, et al. Short group psychoeducation followed by daily electronic self-monitoring in the long-term treatment of bipolar disorders: a multicenter, rater-blind, randomized controlled trial. Int J Bipolar Disord 2019 Nov 4;7(1):23. [doi: 10.1186/s40345-019-0158-8] [Medline: 31680193]
89. Pratap A, Atkins DC, Renn BN, et al. The accuracy of passive phone sensors in predicting daily mood. Depress Anxiety 2019 Jan;36(1):72-81. [doi: 10.1002/da.22822] [Medline: 30129691]
90. Stanislaus S, Faurholt-Jepsen M, Vinberg M, et al. Mood instability in patients with newly diagnosed bipolar disorder, unaffected relatives, and healthy control individuals measured daily using smartphones. J Affect Disord 2020 Jun 15;271:336-344. [doi: 10.1016/j.jad.2020.03.049] [Medline: 32479333]
91. Tseng YC, Lin ECL, Wu CH, Huang HL, Chen PS. Associations among smartphone app-based measurements of mood, sleep and activity in bipolar disorder. Psychiatry Res 2022 Apr;310:114425. [doi: 10.1016/j.psychres.2022.114425] [Medline: 35152069]
92. Turvey C, Fuhrmeister L, Klein D, et al. Secure messaging intervention in patients starting new antidepressant to promote adherence: pilot randomized controlled trial. JMIR Form Res 2023 Dec 8;7:e51277. [doi: 10.2196/51277] [Medline: 38064267]
93. Vachon H, Bourbousson M, Deschamps T, et al. Repeated self-evaluations may involve familiarization: an exploratory study related to ecological momentary assessment designs in patients with major depressive disorder. Psychiatry Res 2016 Nov 30;245:99-104. [doi: 10.1016/j.psychres.2016.08.034] [Medline: 27541343]
94. van den Berg KC, Hendrickson AT, Hales SA, Voncken M, Keijsers GPJ. Comparing the effectiveness of imagery focussed cognitive therapy to group psychoeducation for patients with bipolar disorder: a randomised trial. J Affect Disord 2023 Jan 1;320:691-700. [doi: 10.1016/j.jad.2022.09.160] [Medline: 36206888]
95. Webb CA, Murray L, O Tierney A, Forbes EE. Reward-related predictors of symptom change in behavioral activation therapy for anhedonic adolescents: a multimodal approach. Neuropsychopharmacology 2023;48(4):623-632. [doi: 10.1038/s41386-022-01481-4]

96. van Genugten CR, Schuurmans J, Hoogendoorn AW, et al. Examining the theoretical framework of behavioral activation for major depressive disorder: smartphone-based ecological momentary assessment study. JMIR Ment Health 2021 Dec 6;8(12):e32007. [doi: 10.2196/32007] [Medline: 34874888]
97. Williams MT, Lewthwaite H, Fraysse F, Gajewska A, Ignatavicius J, Ferrar K. Compliance with mobile ecological momentary assessment of self-reported health-related behaviors and psychological constructs in adults: systematic review and meta-analysis. J Med Internet Res 2021 Mar 3;23(3):e17023. [doi: 10.2196/17023] [Medline: 33656451]
98. Jones A, Remmerswaal D, Verveer I, et al. Compliance with ecological momentary assessment protocols in substance users: a meta-analysis. Addiction 2019 Apr;114(4):609-619. [doi: 10.1111/add.14503] [Medline: 30461120]
99. Seidman AJ, George CJ, Kovacs M. Ecological momentary assessment of affect in depression-prone and control samples: survey compliance and affective yield. J Affect Disord 2022 Aug 15;311:63-68. [doi: 10.1016/j.jad.2022.05.015] [Medline: 35537542]
100. Stone AA, Shiffman S. Capturing momentary, self-report data: a proposal for reporting guidelines. Ann Behav Med 2002;24(3):236-243. [doi: 10.1207/S15324796ABM2403_09] [Medline: 12173681]
101. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol 2009;60(1):549-576. [doi: 10.1146/annurev.psych.58.110405.085530] [Medline: 18652544]
102. 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 2024 Feb 27;26:e48168. [doi: 10.2196/48168] [Medline: 38412023]
103. Moon E, Chang JS, Kim MY, et al. Dropout rate and associated factors in patients with bipolar disorders. J Affect Disord 2012 Dec 1;141(1):47-54. [doi: 10.1016/j.jad.2012.02.025] [Medline: 22410504]
104. 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 2020 Feb 15;263:413-419. [doi: 10.1016/j.jad.2019.11.167] [Medline: 31969272]
105. Astill Wright L, Moore M, Reeves S, Perez Vallejos E, Morriss R. Improving the utility, safety, and ethical use of a passive mood-tracking app for people with bipolar disorder using coproduction: qualitative focus group study. JMIR Form Res 2025;9:e65140. [doi: 10.2196/65140] [Medline: 39918865]
106. Gelinas L, Largent EA, Cohen IG, Kornetsky S, Bierer BE, Fernandez Lynch H. A framework for ethical payment to research participants. N Engl J Med 2018 Feb 22;378(8):766-771. [doi: 10.1056/NEJMsb1710591] [Medline: 29466147]
107. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c869. [doi: 10.1136/bmj.c869] [Medline: 20332511]
108. Reeves BC, Gaus W. Guidelines for reporting non-randomised studies. Complement Med Res 2004;11(1):46-52. [doi: 10.1159/000080576]
109. Astill Wright L, Majid M, Shajan G, et al. The user experience of ambulatory assessment and mood monitoring in depression: a systematic review and meta-synthesis. NPJ Digit Med 2025 Dec 2;8(1):737. [doi: 10.1038/s41746-025-02118-8] [Medline: 41331067]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
EMA: ecological momentary assessment
OR : odds ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
RCT: randomized controlled trial

Edited by J Torous; submitted 08.Sep.2025; peer-reviewed by B Najand, M Tusconi; revised version received 27.Nov.2025; accepted 01.Dec.2025; published 12.Jan.2026.

Please cite as:
Astill Wright L, Roe J, Guo B, Morriss R
Dropout, Attrition, Adherence, and Compliance in Mood Monitoring and Ambulatory Assessment Studies for Depression and Bipolar Disorder: Systematic Review and Meta-Analysis
JMIR Ment Health 2026;13:e83765
URL: https://mental.jmir.org/2026/1/e83765
doi:10.2196/83765

Review

Navigating the Digital Landscape for Potential Use of Mental Health Apps in Clinical Practice: Scoping Review

Nikki S Rickard[1], PhD; Perin Kurt[2], MPH; Tanya Meade[2], PhD

[1]Centre for Wellbeing Science, Faculty of Education, The University of Melbourne, Carlton, Australia

[2]School of Psychology, Western Sydney University, Sydney, Australia

Corresponding Author:
Nikki S Rickard, PhD
Centre for Wellbeing Science
Faculty of Education
The University of Melbourne
100 Leicester St
Carlton, 3053
Australia
Phone: 61 400191768
Email: nrickard@unimelb.edu.au

Abstract

Background: The global demand for mental health services has significantly increased over the past decade, exacerbated by the COVID-19 pandemic. Digital resources, particularly smartphone apps, offer a flexible and scalable means of addressing the research-to-practice gap in mental health care. Clinicians play a crucial role in integrating these apps into mental health care, although practitioner-guided digital interventions have traditionally been considered more effective than stand-alone apps.

Objective: This scoping review explored mental health practitioners' views on potential use or integration of smartphone apps into clinical practice. We asked, "What is known about how mental health practitioners view the integration of smartphone apps into their practice?" Further, this scoping review explored the factors that might influence integration of smartphone apps into practice, such as practitioner and client characteristics, app design and functionality, and practitioner views.

Methods: We conducted a systematic search of 3 databases that yielded 38 studies published between 2018 and 2025, involving 1894 participants across various mental health disciplines, most predominantly psychologists and psychiatrists. Data were collected on practitioner and client characteristics, app functionality, and factors deemed important or influencing practitioners' opinions about app integration.

Results: The included studies were most likely to explore use of apps outside the clinical session and focused on self-management apps for mental health monitoring and tracking, and for collecting data from the patient. Fewer studies explored use of apps within-session, or practitioner-guided apps. Practitioners prioritized app features aligned with the American Psychological Association's evaluation criteria, with practitioners prioritizing engagement and interoperability, but also noted the importance of training and resourcing to support integration.

Conclusions: While practitioners recognize the potential of apps in mental health care, integration into clinical practice remains limited. This study highlights the need for further research on practical implementation, clinical effectiveness, and practitioner training to facilitate the transition from potential to actual use of apps in mental health care settings. Recommendations include evaluating effectiveness of app integration through experimental studies and developing training modules to develop practitioners' digital competencies and confidence in app use.

(JMIR Ment Health 2026;13:e75640) doi:10.2196/75640

KEYWORDS

American Psychological Association; anxiety; depression; digital mental health; functionality; mental health practitioner; mobile application; psychiatrist; psychologist; smartphones

Introduction

Background

The global demand for mental health services has increased significantly over the past decade, with the World Health Organization reporting a 13% increase in the period from 2007-2017 [1]. The COVID-19 pandemic has placed an additional burden on the mental health system, with the prevalence of depression and anxiety disorders alone increasing by 25% [2]. Mental health professionals are struggling to meet this increased service demand, and there is a need for sustainable approaches to support professional care [3-5].

Digital resources offer a flexible, data-rich, and economical means of addressing the research-to-practice gap [6-9] and can be used to support the full spectrum of mental health care services [10-13]. Among the most rapidly increasing digital resources available in the mental health sphere are smartphone apps. With over 30,000 mental health smartphone apps (MHapps) available to the public, and smartphone ownership penetrating most sectors of the population, MHapps have the capacity to be an accessible and scalable adjunct to professional mental health care [14-19]. Smartphone apps can also offer quite sophisticated personalization for each user and are reported to be less stigmatizing for many than seeking professional care [17,18,20,21].

Clinicians are in a unique position to guide the integration of smartphone apps into mental health care. The views and capabilities of clinical staff are regarded as critical for successful integration of digital mental health tools [22]. The breadth of their influence is diverse and can include recommending reputable apps to clients to use independently, or using apps in- or out-of-session with ongoing support [23,24]. Digital interventions that are guided by a therapist or with in-person feedback are generally regarded to be more effective than self-management or “stand-alone” apps [16,25-29]. This is consistent with the widely accepted notion of a therapeutic alliance being a key mechanism underlying mental health treatment outcomes [30]. There is, however, some indication that a major benefit of guided digital interventions may be encouraging adherence to the intervention [25,31], and that the role of clinician could be reduced to key touchpoints such as onboarding at the end of every other module or on-demand [25]. While high-quality hybrid models such as the Digital Clinic [32,33] and Precision Behavioral Health [34] offer mobile app functionality (monitoring and interventions) fully integrated with strategically placed practitioner support (eg, through telehealth or in-person follow-ups), there is also a need to better understand how practitioners feel about integrating stand-alone apps into mental health care.

The capacity for smartphone app integration into mental health care also depends on the functionality of the app and the level of support it targets across the mental health spectrum. For example, app functions can include informing, recording, sharing data, reminding, communicating, and displaying [35,36]. Levels of support can range from lower (eg, well-being promotion, mood tracking, assessment, and psychoeducation) to higher (interventions and relapse prevention) intensity [24,37]. Digital mental health services are most confidently advocated as a low-intensity treatment within the stepped-care model, providing health promotion, prevention, and early intervention support for subclinical populations [11,17]. Digital apps are also generally claimed to be more effective for prevention and mild symptomatology [17,38], although some studies have observed stronger effects for individuals with more severe symptoms [39,40]. It would be of interest to better understand where mental health practitioners themselves consider app usage to be most acceptable.

Integration of smartphone apps into mental health care must therefore also consider what factors are important to the practitioner. In a survey of general health care providers, regulatory body support and an evidence base were ranked as the most important factors affecting practitioners’ decisions to incorporate digital tools into their practice [41]. Previous research has also suggested that remote monitoring tools are of particular interest to mental health providers [42]. Several evaluation frameworks describe key criteria identified as important to professional regulatory bodies like the American Psychological Association (APA). The APA’s criteria for evaluating MHapps include accessibility, privacy and security, evidence base, engagement, and interoperability [43,44]. While models such as the APA evaluation framework are critical for shaping regulation of digital tools like smartphone apps, the prioritization of standards may not necessarily align with what is important for individual practitioners. For example, while accessibility is the fundamental standard for MHapps, if practitioners are supplying devices or paying for the cost of the app, then this criterion may become less critical. In a recent evaluation of 100 of the most popular MHapps at the time [37], only one of those apps was found to meet all 5 standards expected by the APA evaluation model. The majority lacked basic accessibility, privacy, and security features, and only one met the final criterion of interoperability, which is important for integrating apps into broader mental health care.

Rationale and Objectives

The use of smartphone apps among physicians and rehabilitation clinicians has previously been explored in other scoping reviews [45,46]; however, to date, no scoping reviews have been conducted on mental health practitioners’ consideration of use or integration of smartphone apps into clinical care. The affordances of smartphone apps may vary between different groups of professionals and in different clinical settings [45]. Accordingly, the aim of this scoping review was to examine research from the past 7 years to answer the broad research question, “What is known about how mental health practitioners view the integration of smartphone apps into their practice?” Further, this scoping review explored the factors that might influence integration of smartphone apps into practice such as practitioner and client characteristics, app design and functionality, and practitioner views.

Methods

Overview

A scoping review methodology was selected to identify and map key characteristics related to the uptake of smartphone apps by mental health practitioners [47]. This review was guided by the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analysis extension for Scoping Reviews; Multimedia Appendix 1 [48]). The protocol was registered through the Open Science Framework [49].

Search Strategy

A search was finalized in March 2025 using 3 electronic databases (PsycINFO, Web of Science, and IEEE Xplore). The search strategy was adapted for each database (Multimedia Appendix 2). Searches were conducted with a combination of terms related to mental health practitioners and smartphone MHapps. These terms were selected following a preliminary review of the literature. Searches were run against the title and abstract, and where possible, subject headings were combined with keywords. The search was restricted to peer-reviewed journal articles written in English and published between 2018 and March 2025.

Eligibility Criteria

Selection of articles was based on the SPIDER (Sample, Phenomenon of Interest, Design, Evaluation, Research Type) framework [50], as shown in Table 1. This framework was chosen to capture qualitative, quantitative, and mixed methods studies.

Table 1. Eligibility criteria.

Category	Inclusion	Exclusion
Sample	• Mental health practitioners, if more than 25% of the sample were psychologists, psychotherapists, psychiatrists, or other counsellors or social workers predominantly working in mental health	• General practitioners, nurses, or other health workers not specific to mental health • Studies that included both practitioners and other end users (eg, clients) but did not differentiate results by participant type
Phenomenon of Interest	• Integration of MHapps[a] into practice. Studies exploring the use, consideration, design, or trial of apps were included	• Studies focusing on web-based mental health apps, as opposed to MHapps • Studies focusing on only users', rather than practitioners', attitudes, behaviors, or experiences toward apps
Design	• Interviews • Surveys • Observations • Behavior measures	• No original data collected
Evaluation	• Attitudes, behaviors or experiences of mental health practitioners in consideration of, or use of, smartphone app integration into their practice	• Studies focusing on only users', rather than practitioners', attitudes, behaviors, or experiences toward apps
Research Type	• Qualitative • Quantitative • Mixed methods	• Reviews and meta-analyses • Commentaries • Opinion pieces • Study protocols

[a]MHapps: mental health smartphone apps.

Selection

Searches were managed using EndNote X9 (Clarivate Analytics) and Microsoft Excel (Microsoft Corp). Duplicates were removed using the EndNote deduplication tool. One reviewer (PK) screened titles and abstracts to exclude obviously irrelevant articles and those identified as reviews, commentaries, or opinion pieces. The remaining titles and abstracts were then screened by 2 independent reviewers (NSR and PK). Disagreements were resolved by discussion, and where doubt existed, an independent assessment was conducted by a third reviewer (TM). Full texts were independently reviewed by all 3 authors. PK reviewed 100% of the articles, with NSR and TM each reviewing approximately 50% of the articles. Disagreements between each pair of reviewers were resolved by discussion to achieve consensus.

Data Extraction and Analysis

An Excel spreadsheet was created to compile relevant data. Data items were sorted into 5 categories:

- Article information: including author, year of publication, country, aim and design.
- Practitioner characteristics: including sample size, type of practitioner (eg, psychologist), age, and gender.
- Client characteristics: including age, mental health condition (eg, depression), and population (eg, clinical).
- Characteristics of app use: including app name, app purpose (prevention/well-being promotion, psychoeducation, monitoring/tracking, assessment/case identification, treatment/intervention, continuing care/relapse prevention, other), app functionality according to the IMS Institute for Healthcare Informatics functionality score (IMS-11; inform, instruct, record [collects, shares, evaluates, intervenes], remind/alert, communicate, display, guide), patterns of use, in- and/or out-of-session, and practitioner guided and/or self-managed.
- Characteristics considered important and/or useful based on APA's criteria: including accessibility, privacy and security, evidence base, engagement, interoperability, and other.

The data charting form was completed by 1 reviewer (PK), and a random subset (20%) of articles was independently verified for accuracy by NSR and TM. Studies were not critically appraised during this review.

Data Synthesis

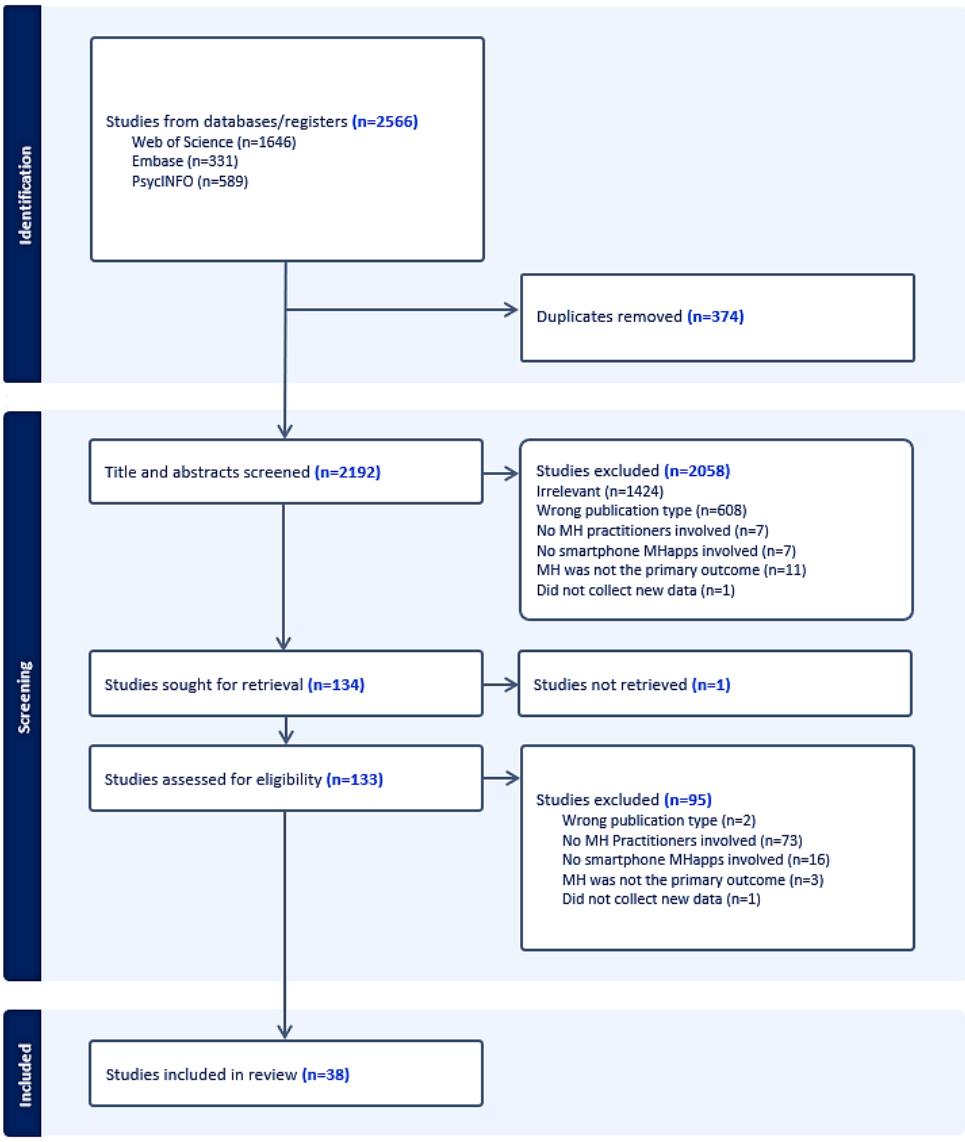
Study designs and methodologies were diverse, so data were synthesized in a descriptive format. Frequencies were determined where possible within variables, and trends were identified where appropriate in a narrative format.

Results

Selection of Sources of Evidence

The initial search yielded 2566 records. After removing 374 duplicates, a total of 2192 titles and abstracts were screened, and 133 full-text publications were reviewed. Following full-text review, 38 publications were submitted to data extraction (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram documenting literature search and outcome of screening process. MH: mental health; MHapps: mental health smartphone apps.



Results of Individual Sources of Evidence and Synthesis of Results

The full characteristics of the included studies are outlined in Tables 2 and 3, with detailed information provided in Table S1 in Multimedia Appendix 3.

Table 2. Study and participant information of retrieved studies.

Author	Year	Country	Primary methodology	Sample size meeting criteria, n (% of N)	Practitioner age[a] (years), mean (SD) or mode (SD)	Client age (years), mean (SD) or median (range)
Adams et al [51]	2021	US[b]	Mixed methods	11 (55% of 20)	• Alpha: 43.3 (7.5) • Beta 2: 37.8 (12.1)	• Alpha: 15.3 (1) • Beta: 15.9 (1)
Almadani et al [52]	2025	Saudi Arabia	Quantitative	135 (35% of 386)	• <30 (61.1%) • 30-40 (30.1%) • 41-50 (06.0%) • >50 (02.8%)	• Not reported
Anastasiadou et al [53]	2019	Spain	Qualitative	7 (88% of 8)	• 34.63 (7.21)	• 15 (0.5)
Armstrong et al [54]	2021	US	Qualitative	6 (50% of 12)	• 38.42 (5.78)	• 18.9 (3.73)
Bucci et al [55]	2019	UK[c]	Qualitative	17 (35% of 48)	• 36.2 (SD not reported)	• Not reported
Chang [56]	2023	US	Qualitative	5 (100% of 5)	• Not reported	• 37 (10.7)
Cheung et al [57]	2023	Canada	Quantitative	98 (91% of 108)	• Mode: 60+ (34.3%)	• 18-29 (20.5%) • 30-39 (24.4%) • 40-49 (20.6%) • 50-59 (22.9%) • 60+ (13%)
Deady et al [58]	2023	Australia	Mixed methods	3 (100% of 3)	• Not reported	• 50 (range 42-56)
Dobson et al [59]	2022	New Zealand	Mixed methods	178 (91% of 195)	• Mode: 24-35 (37%)	• Not reported
Dominiak et al [60]	2024	Poland	Mixed methods	62 (42% of 148)	• 25-39 (27%) • 40-55 (56%) • 55-64 (15%)	• Not reported
Dubad et al [61]	2021	UK	Mixed methods	3 (50% of 6)	• Not reported	• 20.71 (2.56)
Etingen et al [62]	2024	US	Mixed methods	230 (81% of 284)	• Under 35 (19.2%) • 36-45 (50.5%) • 46-55 (19.2%) • 56-65 (9.3%) • 66-75 (1.6%)	• Not reported
Francese et al [63]	2023	Italy	Quantitative	8 (100% of 8)	• 43 (6.18)	• 36.24 (14.02)
Gonzalez-Perez et al [64]	2024	Spain	Mixed methods	9 (100% of 9)	• 35 (SD not reported)	• Not reported
Green et al [65]	2023	US	Mixed methods	7 (78% of 9)	• Not reported	• 19.6 (2.05)
Heydarian et al [66]	2023	Iran	Mixed methods	7 (54% of 13)	• Not reported	• Not reported
Hildebrand et al [67]	2024	Germany	Qualitative	269 (77% of 350)	• 42.83 (12.16)	• Not reported
Hoffman et al [68]	2019	US	Mixed methods	15 (63% of 24)	• Not reported	• 36.5
Kerst et al [69]	2020	Germany	Quantitative	33 (58% of 57)	• 43 (12.3)	• Not reported

Author	Year	Country	Primary methodology	Sample size meeting criteria, n (% of N)	Practitioner age[a] (years), mean (SD) or mode (SD)	Client age (years), mean (SD) or median (range)
Khan et al [70]	2023	Canada	Qualitative	2 (33% of 6)	• Range: 32-48	• Range: 32-48
Li et al [71]	2022	Australia	Quantitative	10 (31% of 32)	• 38.7 (10.4)	• 14.94 (1.3)
Lukka et al [72]	2023	Finland	Qualitative	11 (61% of 19)	• 18-29 (5%) • 30-39 (16%) • 40-49 (32%) • 50-59 (37%) • 60-69 (11%)	• Not reported
McGee-Vincent et al [73]	2023	US	Quantitative	271 (25% of 1107)	• MHSL staff (44.2, 10.2) • AOSL staff (46.0, 10.9)	• Not reported
Medich et al [74]	2023	US	Qualitative	10 (63% of 16)	• Not reported	• Median: 45, Range: 21-66 years
Miller et al [75]	2019	US	Mixed methods	103 (47% of 220)	• 45.5 (11.1)	• Not reported
Morton et al [76]	2021	Multiple	Mixed methods	57 (71% of 80)	• 44.7 (13.1)	• Not reported
Naccache et al [77]	2021	France	Mixed methods	3 (43% of 7)	• 36.7 (7.38)	• 15.5 (1.07)
Nogueira-Leite et al [78]	2023	Portugal	Mixed methods	152 (95% of 160)	• <26 (2.5%) • 26-35 (35%) • 36-45 (36.9%) • 46-55 (19.4%) • 56-65 (3.8%) • >65 (2.5%)	• Not reported
Orengo-Aguayo et al [79]	2018	US	Mixed methods	7 (78% of 9; Phase 1) and 35 (63% of 56; Phase 2)	• Phase 1: 41.88 (10.25) • Phase 2: 39.73 (10.26)	• Not reported (although we know that the clinicians treated children and adolescents)
Patoz et al [80]	2021	France	Qualitative	16 (62% of 26)	• 45.5 (12.2)	• 51.5 (15.5)
Puhy et al [81]	2021	US	Mixed methods	3 (100% of 3)	• Not reported	• 16.1 (1.2)
Richards et al [82]	2018	Australia	Qualitative	6 (100% of 6)	• Not reported	• Not reported (although we know that the patients were 18 years and older)
Rodriguez-Villa et al [83]	2021	Multiple	Qualitative	35 (67% of 52)	• BIDMC[d]: 41 (SD not reported) • AIIMS[e]: 35 (SD not reported) • NIMHANS[f]: 41 (SD not reported)	• BIDMC: 32 (SD not reported) • AIIMS: 33 (SD not reported) • NIMHANS: 35 (SD not reported)
Rothmann et al [84]	2022	Denmark	Qualitative	3 (100% of 3)	• Not reported	• Only reported range = 25-59 years
Stefancic et al [85]	2022	US	Qualitative	8 (73% of 11)	• Not reported	• Not reported (although we know the clinics provide treatment to adolescents and young adults aged 16-30 years)
Strodl et al [86]	2020	Australia	Mixed methods	38 (60% of 63)	• Focus group: 47 (10.5) • Telephone sample: 45.7 (10.5)	• Not reported

Author	Year	Country	Primary methodology	Sample size meeting criteria, n (% of N)	Practitioner age[a] (years), mean (SD) or mode (SD)	Client age (years), mean (SD) or median (range)
Weermeijer et al [87]	2023	Belgium	Mixed methods	9 (75% of 12)	• Users: 45.57 (6.11) • Dropouts or nonusers: 43.50 (17.50)	• Users: 34.93 (11.27) • Dropouts or nonusers: 36.67 (13.47)
Wu et al [88]	2020	US	Mixed methods	12 (100% of 12)	• Not reported	• Not reported

[a]Mean age and gender typically reported for all practitioners, and sometimes all of sample.

[b]US: United States.

[c]UK: United Kingdom.

[d]BIDMC: Beth Israel Deaconess Medical Center.

[e]AIIMS: All India Institute of Medical Sciences.

[f]NIMHANS: National Institute of Mental Health and Neurosciences.

Table 3. Characteristics of app use in retrieved studies, and American Psychiatric Association (APA) standards prioritized by practitioners.

Author	Use of app for research or practice	App purpose	In-session and/or out-of-session	Practitioner guided and/or self-managed	APA standards prioritized
Adams et al [51]	Research	• Prevention/promotion • Assessment/case identification	Both	Both	A[a], SP[b], EB[c], E[d], I[e]
Almadani et al [52]	Clinical	• Psychoeducation monitoring/tracking • Treatment/intervention	Not reported	Not reported	A, SP, EB
Anastasiadou et al [53]	Research	• Monitoring/tracking • Continuing care/relapse prevention	Out-of-session	Practitioner guided	A, SP, EB, E, I
Armstrong et al [54]	Research	• Monitoring/tracking	Out-of-session	Not reported	A, SP, EB, E, I
Bucci et al [55]	Research	• Monitoring/tracking • Treatment/intervention	Out-of-session	Self-managed	A, SP, E, I
Chang et al [56]	Clinical	• Monitoring/tracking	Out-of-session	Not reported	A, SP, EB
Cheung et al [57]	Clinical	• Assessment/case identification • Monitoring/tracking	Not reported	Not reported	A, SP
Deady et al [58]	Research	• Psychoeducation • Treatment/intervention	Out-of-session	Practitioner guided	A, EB, E, I
Dobson et al [59]	Clinical	• Not reported	Not reported	Not reported	SP
Dominiak et al [60]	Clinical	• Prevention/well-being promotion • Monitoring/tracking • Continuing care/relapse prevention	Not reported	Not reported	A, EB, E, I
Dubad et al [61]	Research	• Monitoring/tracking	Not reported	Not reported	I
Etingen et al [62]	Research	• Monitoring/tracking • Assessment/case identification	Out-of-session	Practitioner guided	E, I
Francese et al [63]	Research	• Assessment/case identification • Monitoring/tracking	In-session	Practitioner guided	E, I
Gonzalez-Perez et al [64]	Research	• Psychoeducation • Monitoring/tracking • Assessment/case identification • Treatment/intervention	Out-of-session	Self-managed	Not available
Green et al [65]	Clinical	• Psychoeducation • Monitoring/tracking • Treatment/intervention	Out-of-session	Practitioner guided	A, EB, E, I
Heydarian et al [66]	Research	• Psychoeducation • Monitoring/tracking	Not reported	Self-managed	SP, EB, I
Hildebrand et al [67]	Clinical	• Psychoeducation • Treatment/intervention	Out-of-session	Not reported	A, EB, E
Hoffman et al [68]	Clinical	• Treatment/intervention • Psychoeducation • Monitoring/tracking	Out-of-session	Self-managed	A, SP, EB, E, I
Kerst et al [69]	Clinical	• Treatment/intervention	Not reported	Not reported	A, SP, I
Khan et al [70]	Clinical	• Psychoeducation • Monitoring/tracking • Treatment/intervention	Not reported	Not reported	SP, E, I

Author	Use of app for research or practice	App purpose	In-session and/or out-of-session	Practitioner guided and/or self-managed	APA standards prioritized
Li et al [71]	Research	• Treatment/intervention (cognitive behavioral therapy) • Psychoeducation (activities for depression)	Out-of-session	Self-managed	A, SP, EB, E, I
Lukka et al [72]	Clinical	• Psychoeducation • Treatment/intervention	Not reported	Both	SP, EB, E
McGee-Vincent et al [73]	Clinical	• Not reported	Not reported	Not reported	A, SP, EB, E
Medich et al [74]	Research	• Monitoring/tracking	Out-of-session	Self-managed	EB, E
Miller et al [75]	Clinical	• Monitoring/tracking	Out-of-session	Self-managed	A, SP, EB, E, I
Morton et al [76]	Clinical	• Monitoring/tracking • Treatment/intervention	Not reported	Self-managed	A, SP, EB, E, I
Naccache et al [77]	Research	• Prevention/promotion • Monitoring/tracking • Treatment/intervention	Not reported	Self-managed	A, SP, EB, E, I
Nogueira-Leite et al [78]	Clinical	• Monitoring/tracking	Not reported	Not reported	A, EB, E, I
Orengo-Aguayo et al [79]	Research	• Prevention/promotion • Monitoring/tracking • Assessment/case identification	Both	Practitioner guided	A, SP, EB, E, I
Patoz et al [80]	Research	• Prevention/promotion • Monitoring/tracking • Assessment/case identification	Out-of-session	Self-managed	A, SP, EB, E, I
Puhy et al [81]	Research	• Prevention/promotion • Monitoring/tracking	Both	Both	SP, EB, E, I
Richards et al [82]	Research	• Tracking/monitoring • Assessment/case identification	Out-of-session	Both	A, E, I
Rodriguez-Villa et al [83]	Research	• Continuing care/relapse prevention	Both	Self-managed	A, SP, EB, I
Rothmann et al [84]	Research	• Prevention/promotion • Treatment/intervention	Both	Practitioner guided	A, E
Stefancic et al [85]	Research	• Assessment/case identification	Out-of-session	Self-managed	A, SP, EB, E, I
Strodl et al [86]	Research	• Prevention/promotion • Tracking/monitoring • Assessment/case identification	Out-of-session	Both	A, SP, EB, E, I
Weermeijer et al [87]	Research	• Monitoring/tracking	In-session	Not reported	A, EB, E
Wu et al [88]	Clinical	• Monitoring/tracking	Out-of-session	Practitioner guided	EB, E, I

[a]A: accessibility.

[b]SP: security and privacy.

[c]EB: evidence base.

[d]E: engagement.

[e]I: interoperability.

Studies' Characteristics

Publications were distributed evenly across the 8-year period of inclusion (2018-2025), with a peak in 2023 during which 32% (12/38) of included studies were published (Table 2). The largest representation of studies was from the United States (12/38, 32% of studies), followed by Australia (4/38, 11% of studies), with only one or two studies across each of the remaining 14 countries. The total number of practitioners sampled within the retrieved studies was 1894, with sample sizes ranging from 3 [84] to 271 [73], and 11 being the median sample size.

The studies' aims (Table S1 in Multimedia Appendix 3) covered a wide range of topics related to the use of mobile apps in mental health care, with about half of the retrieved studies reporting on clinicians' attitudes, experiences, and their views on acceptability and feasibility of using digital tools in mental health care. Usability studies involved developing and testing MHapps for a variety of conditions, with depression the most common, followed by general mental health and multiple disorders. Less commonly, studies focused on evaluating specific features of mental health apps, such as mood monitoring components, or on developing training programs for health care professionals and assessing their digital competence. The studies used a diverse range of methodological approaches, with the majority using qualitative or mixed methodologies to explore the perspectives, experiences, and needs of stakeholders regarding the use of mobile apps in mental health care. More than half the studies (20/38, 53% of total) were mixed methods designs, while 12 studies (32% of total) were qualitative designs (involving for example, focus groups, semistructured interviews). Only 6 of the 38 (16%) studies had an entirely quantitative design (involving cross-sectional and longitudinal surveys or questionnaires). Several studies focused in designing new apps and included co-design or user-centered design approaches, which emphasized the involvement of end-users (practitioners, clients, or both) in the development and refinement of mobile apps.

Practitioner and Client Information

A range of mental health practitioners were represented in the included studies (Table S1 in Multimedia Appendix 3), with psychologists the most common (in 32 studies) and psychiatrists the next most common (in 21 studies). Practitioner age varied significantly across studies, with modal age ranging from <30 years to >60 years, and there were more female (40.7% to 100% range) than male or nonbinary practitioners represented. Clients' age ranged from 15-52 years, with the most common age range (when reported) being adolescents and young adults (8 of the 38 studies). The majority of studies (29 of 38 studies) appeared to include clinical samples, although reporting was at times ambiguous. The most commonly reported mental health conditions in the client sample were mood disorders (depression, anxiety, and bipolar disorder), following by general mental health conditions and multiple disorders.

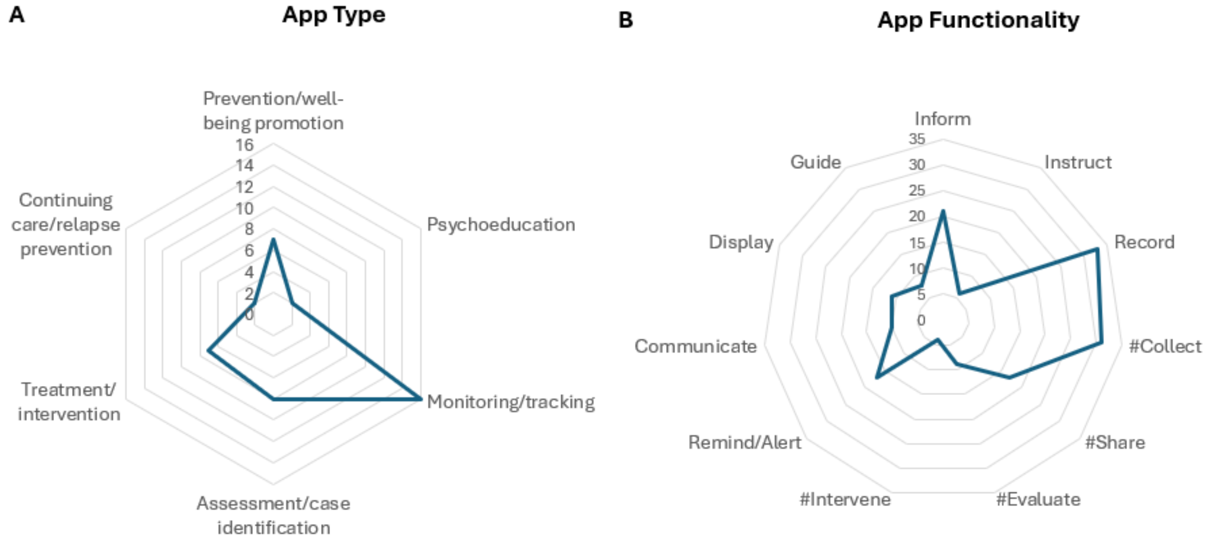
App Characteristics

Less than half of the included studies (16/38, 42%) involved practitioners' use of apps integrated into their routine clinical practice, rather than as part of trials or feasibility research studies (Table 3). Almost half of the studies (18/38, 47%) explored the use of a specific app, with 4 (11%) studies examining a toolkit of apps (Table S1 in Multimedia Appendix 3). The remaining studies explored practitioners' attitudes or use of mental health apps more generally. The majority of apps discussed across studies were monitoring/tracking apps, followed by treatment/intervention apps, psychoeducation, and assessment/case identification apps (Table 3 and Figure 2A). With regard to app functionality (Figure 2B), the majority (25/38, 66%) of studies cited apps which had 3 or fewer of the 7 IMS-11 app functions. Most studies (33/38, 87%) reported on apps that included a recording function, with nearly all studies (31/38, 82%) reporting on apps that collected data from users. The next most common function reported in studies was informing (21/38, 55%), followed by reminding/alerting and recording for the purpose of sharing data (each reported in 17/38, 45% of studies). The least common functions were instructing (reported in only 6/38, 16% of studies) and intervening (reported in 4/38, 11% of studies).

Apps were used primarily out-of-session (18/38, 47% of studies), with 5 studies (13%) citing app use both in- and out-of-session. Only 2 studies [63,87] focused on an app used specifically in clinical sessions. The most common management approach reported was self-management (12/38, 32% of studies), with 5 studies (13%) reporting both practitioner-guided and self-management, and 8 studies (21%) reporting practitioner-guided only. The remainder of studies did not report the management approach.

A total of 16 of the 38 (42%) retrieved studies indicated practitioners' current patterns of using apps with their clients. In the majority of these studies (31/38, 82%), practitioners had incorporated apps into their practice by either using them or recommending them to clients. Patterns of use included providing a list of apps to patients to explore, or recommending a specific app with varying levels of instructions or follow up. Where described, practitioner's intentions tended to be for patients to self-manage their health [68,75,76,83], with follow-up or review rarely reported. In other studies, practitioners intended to use apps in the future for their patients [52,78]. In one study, the likelihood of prescribing apps increased following practitioner training designed to increase the reach of MHapps for veterans [73].

Figure 2. Radar plots of (A) purpose of apps and (B) app functionality (IMS Institute for Healthcare Informatics functionality score [IMS-11]). Categories prefixed with # represent the 4 subfunctions of the recording function.



Practitioners' Prioritization of App Features Mapped Onto APA App Quality Criteria

All 5 APA criteria were recognized as important across the majority of studies (Table 3). The criterion most commonly raised was engagement, cited by practitioners in 79% (30/38) of retrieved studies. Factors of importance to practitioners included ease of use and user-friendliness, attractive and intuitive design, ability to personalize and customize features, inclusion of reminders and notifications, gamification and interactive elements, and provision of feedback and progress tracking. For example, Li et al [71] emphasized the importance of incorporating interaction patterns and functionalities used in popular social media apps, such as swipe interactions and short videos, to enhance engagement for a youth-focused app.

Similar in importance to practitioners was Interoperability, cited in 76% (29/38) of studies, with key factors including the ability to share data with clinicians or caregivers, integration with electronic health records, supporting in-session and between-session use, and compatibility with clinician workflows and systems. For example, Wu et al [88] highlighted the importance of apps that seamlessly integrate data into electronic health records and support collaborative care, suggesting the value of interoperability for practitioners.

Accessibility and evidence base were also commonly cited criteria by practitioners (cited in 26/38, 68% and 27/38, 71% of the studies respectively). Accessibility included factors associated with smartphone ownership (due to technical resources such as Wi-Fi or cost), availability on multiple platforms, language requirements, ability to use offline or on demand, and data storage requirements. Regarding the importance of an evidence base, practitioners noted the importance of including credible, reliable, and up-to-date content; the use of validated scales and questionnaires; alignment with clinical practice guidelines; provision of psychoeducation; information being appropriate for specific mental health conditions; and information about the underlying science and research. For example, Miller et al [75] noted that the intention to use apps in the future was influenced by professional training, followed by scientific evidence, suggesting the importance of an evidence base for practitioners. Security and privacy concerns were the least frequently cited issue by practitioners although still cited in 63% (24/ 38) of studies. Concerns included data privacy and confidentiality, security protocols (eg, passwords, encryption), regulations and liability governing data access and use, and ensuring user privacy when sharing information with clinicians or caregivers. For example, Anastasiadou et al [53] highlighted concerns about a restrictive health care system in Spain that limited the sharing of patient information online, emphasizing the importance of addressing privacy and security concerns. Given all 5 APA criteria were raised as important across most studies, it was not possible to discern any patterns or associations between APA criteria raised and client or practitioner characteristics.

Other issues identified as important by practitioners beyond the formal APA app quality evaluation model included a range of implementation factors, such as the importance of clinician training in using the technology to enhance familiarity and confidence in using and recommending apps, and organizational investment in resources and support for integrating digital mental health into practice. Client factors were also noted, including the potential impact of apps on therapeutic alliance and treatment outcomes, and moderation of app utility by client characteristics, including digital literacy, age, and motivation.

Discussion

Overview

The aim of this scoping review was to explore what is known about the mental health practitioners' views on integration of smartphone apps into their practice; the factors that may influence such integration, including practitioner and client characteristics, app functionality and design; and reported use of those apps in practice.

Main Findings

Across the 38 included studies, a total of 1894 mental health professionals participated, with sample sizes ranging from 3 to 271. The participants predominantly included psychologists and psychiatrists, as well as a number of other mental health practitioners, with participants being predominantly female and mid-career based on their age range (30-39 years and 40-49 years). Practitioners worked across a range of services, from specialized clinics (eg, eating disorders and psychosis early intervention) to school- or tertiary-based counselling units, demonstrating a diverse and representative sample of mental health practitioners and services. Notably, however, more than half (24 of the 38) of the studies had sample sizes less than 50, and some demographic features such as age or gender, were not reported by all studies, which limits the degree of generalizability of views to a broader mental health practitioner population.

From the demographic information included, the participants appear to be predominantly experienced and qualified practitioners actively involved in mental health services. Practitioners' clients across the included studies ranged from children to adolescents and adults, predominantly from clinical settings and presenting with a range of mental health issues, most commonly mood and anxiety disorders. This is consistent with the prevalence of those mental health conditions in both clinical and community populations, which have further increased post the COVID-19 pandemic [2]. The increased prevalence of those common conditions continues to challenge limited health services, access and continuity of care and practitioners' ability to meet the needs of growing number of clients. The consistency in the mental health conditions in the population, and those identified in this review of practitioners who are exploring smartphone apps in their practice, is therefore promising for alignment of support demand and practitioner readiness for integration of digital tools into their practice. Therefore, smartphone apps offer a viable add-on resource and a self-directed mental health support.

However, much of the current literature on the use of smartphone apps in mental health care is predominantly focused on the development of such apps and exploration of perspectives of both practitioners and clients on their usability rather than on actual usage patterns or clinical effectiveness. This aligns with recognition of the importance of co-designing apps to engage end users (which can be both practitioners and clients), and better understand their needs and preferences [22]. Across the 38 studies, about half reported app use out-of-sessions, or in- and out-of-session, with only 2 out of 38 reporting in-session use only. It therefore appears that while the use of apps is recognized by practitioners as potentially helpful to their clients, the actual use remains consistent with stepped-care models of mental health care, in which digital technologies are regarded as best suited for self-managed, low-intensity support [12,89] rather than as part of an integrated in-session care plan. This is understandable, given that the primary focus of the apps included across those studies was on well-being promotion, tracking/monitoring, psychoeducation, early intervention, and assessment monitoring, as well as their capacity to collect, record, and inform. In contrast, about one-third of studies reported using apps within the higher-intensity range of treatment/intervention or continued care/relapse prevention functions, with few apps offering "instructing" or "intervening" functions. Interestingly, however, the proportion of apps offering these higher-intensity support options appears to be increasing, with the proportion doubling from 17% to 34% in the final year covered by this review. Most apps remain limited in the scope of their functionality, offering 3 or fewer functions. Apps that offer a greater multifunctionality in their design are more likely to be useful to client-practitioner sessions [36], suggesting a need for broader spectrum, "all-in-one" apps.

Despite previous literature suggesting practitioner-guided apps are more effective than self-managed or "stand-alone" apps [16,25), the studies identified in this review focused more on self-management apps. The majority of studies (17/38, 45%) included self-management apps, with 32% (12/38) entirely stand-alone. While there is currently still limited evidence of effectiveness for stand-alone apps [16], this delivery mode may be of greater interest to practitioners given the persistent high demand for their services [90]. Self-management apps also align with models of health care that aim to empower patients to participate more in the management of their own health [91]. This is despite some reports that involvement of practitioners in client use of mental health apps aids effectiveness of those apps in comparison to self-managed or stand-alone app use [26,27]. There is, however, a need for research to capture what that involvement looks like and to what degree it facilitates the effectiveness of the app and its contribution to the clinical sessions. Practitioners are more likely to recommend apps (whether used in- or out-of-session) if there is clear evidence of how their involvement and the client's use of the app contributes to the clinical care outcomes.

Given that the majority of mental health apps do not meet all 5 standards outlined by the APA evaluation model [37], it is understandable that their use by mental health practitioners is still at a potential rather than actual and evidence-based stage. Notably, across the 38 studies reviewed, practitioners raised the importance of app features consistent with those 5 APA standards, indicating an informed position in their considerations of suitability of apps for integration into mental health care. It is notable that engagement and interoperability were most commonly identified by practitioners as important. Practitioners' awareness of the need for an engaging user experience is promising, given its critical role in compliance and sustainability of use [9,92,93]. Similarly, the importance of data sharing is consistent with digital tools being used as an adjunct rather than alternative to professional care, and providing practitioners with information that may contribute to their in-session interactions with the client [44,94]. Integration of assessment and monitoring data into electronic records may also provide new insights which may not be possible with more manual recording of less regular data points. It is also promising that an evidence or clinical base for apps was reported in 71% (27/38) of studies. This reflects an awareness that mental health apps need to be credible and developed by trusted sources, and for the function and outcomes of the app to be aligned through systematic evaluation. In this context, it is of concern that the majority of mental health apps available publicly lack sufficient evidence or clinical support for their effectiveness [15,95].

Security and privacy and accessibility were each also reported across the majority of studies. The importance placed on security and privacy demonstrates practitioners' awareness of the importance of ethical management of digital patient records [96]. The equally high importance of accessibility reflects practitioners' awareness that mental health apps must be designed and distributed in a way that does not exclude any sociodemographic groups, and encourages uptake and sustained use for all. Among the other criteria raised by practitioners beyond the APA model, the need for organizational support in training and technical resource support was a commonly reported concern. Implementation of training modules has been effectively demonstrated to enhance knowledge and confidence of mental health practitioners about using apps [73]. However, training on digital mental health options for practitioners remains limited [97]. These are important practical considerations in facilitating progression from potential to actual use of apps in clinical settings and should be considered for all mental health staff working within clinical settings.

To progress the integration of mental health apps into clinical sessions, it may be informative to understand the context of clinical sessions that include assessment, formulation, treatment plan, monitoring of progress, and self-directed tasks between sessions. In addition, it would be helpful to include information on which APA standards a mental health app meets in app stores, or some other mental health register, to assist selection for a particular client with a particular mental health condition. Without a closer alignment between what may aid sessions and a client's in- and out-of-session support, self-development- and psychoeducation, there may be too many apps for practitioners to choose from without a clear sense of how useful those apps may be or how to measure their add-on effectiveness.

Overall, the 38 studies included in this review provide an informative understanding of what practitioners consider important in the design and function of mental health apps if they were to use or integrate their use into their clinical sessions. Further research is required to capture what that use may look like in practice and whether using particular apps provides additional benefit to the clinical session and the therapeutic output, or provides support in between sessions that may otherwise be lacking. Given the constrains of limited mental health services, anything that may be added to the sessions' content (ie, review of app-collected data) or between sessions (ie, client and practitioner's engagement with the app) must have a clear add-on clinical value.

Limitations

Limitations across studies include small sample sizes, limited demographic and descriptive information, overrepresentation of studies from the United States and early adopters of integration of apps into clinical practice, which means that the findings across those studies may not be generalizable to broader populations of practitioners. Further, the studies included in this scoping review were inclusive of studies exploring both views of potential and actual integration of apps into practice and therefore cannot provide indication of prevalence of use. While some of the studies have provided indication of factors that may influence integration, or not, of apps into clinical practice, this is an area that requires further research across broader practitioners' representation. Further, real-world factors relating to successful integration of apps into practice—such as patient acceptability, organizational culture around adoption, and resourcing and feasibility of sustained use—were not within the scope of this review [22,98].

This review also focused on mental health professionals only to contain the scope of practice. A similar review could be undertaken with other health professionals who work with mental health issues in a different therapeutic context (eg, general practitioners, nurses, and other allied health practitioners), which may reveal different uses of available apps. The quality of included studies was also not assessed, as this was a scoping, not a systematic, review and therefore relied on descriptive rather than evaluative presentation of those studies.

Future Directions and Recommendations

Notwithstanding these limitations, the 38 studies offer an overview of several useful mental health apps that could be integrated into clinical practice, ranging from between-session monitoring/tracking data to interactive relapse prevention apps that may provide useful between-session engagement and within-session progress-informing data. This scoping review has identified that practitioners are interested in the use of low-intensity support apps outside of sessions, which is aligned with the stepped-care model and 2 key points: addressing support gaps between sessions, which are critical to retention and relapse prevention; and empowering clients to self-manage some aspects of their health.

As a way forward, it is recommended that the potential use of some apps be explored in a clinical setting using well-structured experimental designs to assess suitability, usability, and effectiveness of those apps beyond their potential and ad hoc use. A good starting point may be integrating the use of apps outside of sessions, with a review of their use, clients' perceptions of usefulness, and, where relevant, review of data during sessions, similar to the common practice of giving clients homework (eg, new skills to practice, daily mood-monitor sheets, or behavior tracking).

For practitioners who would like to use mental health apps, it is recommended that they base their selection of apps on the APA standards as a starting point, focusing on apps that align with particular aspects of a client's care plan (eg, relaxation for anxiety or daily pleasant activities for mood), and then determine how they may like to integrate them in their clinic sessions (eg, for psychoeducation or monitoring symptoms). Some practitioners may find reviewing apps against the APA standards relatively straightforward, especially if they are comfortable users of digital technology. Others, however, may require some support, given that there is a large variation in digital competencies, with many having low awareness and knowledge [59,68,69] and confidence in suggesting apps to their clients [76]. It is therefore recommended that educational training modules or workshops are developed to inform use of mental health apps in practice, starting with low-intensity app selection for use between session to high-intensity apps for within sessions for more advanced digital technology-user practitioners. Such training needs to address commonly raised concerns by both practitioners and clients, including accessibility, safety, and potential side effects [80], and build practitioners' self-efficacy in the use of apps with their clients [76]. An implementation science lens [22,99] would be useful to extend this review into consideration of factors that support successful integration of mental health apps into practice.

Conclusion

This scoping review has provided valuable insights into mental health practitioners' perspectives on integrating smartphone apps into their practice. The findings reveal that practitioners are interested in the capabilities of these digital tools, particularly for self-management and out-of-session support, aligning with stepped-care models of mental health care. The use of mental health apps is an untapped resource that can support both practitioners and clients in improving engagement and implementation of care plan strategies [86] and a reciprocal responsibility for mental health management [82]. This may be particularly relevant for practitioners who are regular uses of digital technologies and who can be at the forefront of apps integration into practice, as well as practitioners whose clients are highly engaged with technology as a health self-management resource. However, the actual integration of apps into clinical practice remains limited, with practitioners emphasizing the importance of app features such as interoperability, security, privacy, engagement, and accessibility. This review highlights a gap between consideration and actual use of mental health apps in clinical settings, suggesting a need for further research to explore the practical implementation and effectiveness of these tools in therapeutic contexts. To facilitate the adoption of mental health apps in clinical practice, it is recommended that:

- Practitioners base their app selection on established standards, such as those outlined by the APA.
- Educational training modules or workshops be developed to enhance practitioners' digital competencies and confidence in using apps with clients.
- Well-structured experimental designs be used to assess the suitability, usability, and effectiveness of apps in clinical settings.
- App developers and researchers work toward aligning app functionalities more fully with a range of therapeutic needs and contexts.

As the field of digital mental health continues to evolve, ongoing collaboration between practitioners, researchers, and app developers will be crucial in realizing the full potential of smartphone apps as valuable adjuncts to traditional mental health care.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR checklist.
[PDF File (Adobe PDF File), 499 KB - mental_v13i1e75640_app1.pdf]

Multimedia Appendix 2
Search strategy detail.
[DOCX File , 24 KB - mental_v13i1e75640_app2.docx]

Multimedia Appendix 3
Additional characteristics of studies, sample and apps in retrieved studies.
[DOCX File , 44 KB - mental_v13i1e75640_app3.docx]

References

1. World Mental Health Report: Transforming Mental Health for All. Geneva: World Health Organization; 2022.
2. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021;398(10312):1700-1712 [FREE Full text] [doi: 10.1016/S0140-6736(21)02143-7] [Medline: 34634250]
3. Holmes EA, Ghaderi A, Harmer CJ, Ramchandani PG, Cuijpers P, Morrison AP, et al. The lancet psychiatry commission on psychological treatments research in tomorrow's science. Lancet Psychiatry 2018;5(3):237-286. [doi: 10.1016/S2215-0366(17)30513-8] [Medline: 29482764]
4. Kazdin AE, Blase SL. Rebooting psychotherapy research and practice to reduce the burden of mental illness. Perspect Psychol Sci 2011;6(1):21-37. [doi: 10.1177/1745691610393527] [Medline: 26162113]
5. Wainberg ML, Scorza P, Shultz JM, Helpman L, Mootz JJ, Johnson KA, et al. Challenges and opportunities in global mental health: a research-to-practice perspective. Curr Psychiatry Rep 2017;19(5):28 [FREE Full text] [doi: 10.1007/s11920-017-0780-z] [Medline: 28425023]

6. Scoping and development of a National Digital Mental Health Framework: Current State Assessment Report.: Australian Government Department of Health; 2020. URL: https://www.pwc.com.au/health/National-Digital-Mental-Health-Framework_Current-State-Assessment-Report.pdf [accessed 2025-03-23]
7. East ML, Havard BC. Mental health mobile apps: from infusion to diffusion in the mental health social system. JMIR Ment Health 2015;2(1):e10. [doi: 10.2196/mental.3954] [Medline: 26543907]
8. McGorry PD, Mei C, Chanen A, Hodges C, Alvarez-Jimenez M, Killackey E. Designing and scaling up integrated youth mental health care. World Psychiatry 2022;21(1):61-76 [FREE Full text] [doi: 10.1002/wps.20938] [Medline: 35015367]
9. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. JMIR Ment Health 2020;7(3):e18848 [FREE Full text] [doi: 10.2196/18848] [Medline: 32213476]
10. Cornish P. Stepped Care 2.0: A Paradigm Shift in Mental Health. Cham: Springer; 2020.
11. PHN Mental Health Flexible Funding Pool Programme Guidance: Stepped Care.: Australian Government Department of Health; 2016. URL: https://www.health.gov.au/sites/default/files/documents/2021/04/primary-health-networks-phn-primary-mental-health-care-guidance-stepped-care.pdf [accessed 2025-03-23]
12. Pedrelli P, Bentley K, Pittman M, Meyer A, Fisher L. Use of technology and stepped care. In: Current Clinical Psychiatry. Cham: Springer International Publishing; 2023:411-422.
13. Rickwood D. Entering the e-spectrum: an examination of new interventions for youth mental health. Youth Studies Australia 2012;34(4):18-27 [FREE Full text]
14. De Witte NAJ, Joris S, Van Assche E, Van Daele T. Technological and digital interventions for mental health and wellbeing: an overview of systematic reviews. Front Digit Health 2021;3:754337 [FREE Full text] [doi: 10.3389/fdgth.2021.754337] [Medline: 35005695]
15. Lecomte T, Potvin S, Corbière M, Guay S, Samson C, Cloutier B, et al. Mobile apps for mental health issues: meta-review of meta-analyses. JMIR Mhealth Uhealth 2020;8(5):e17458 [FREE Full text] [doi: 10.2196/17458] [Medline: 32348289]
16. Moshe I, Terhorst Y, Philippi P, Domhardt M, Cuijpers P, Cristea I, et al. Digital interventions for the treatment of depression: a meta-analytic review. Psychol Bull 2021;147(8):749-786. [doi: 10.1037/bul0000334] [Medline: 34898233]
17. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017;16(3):287-298 [FREE Full text] [doi: 10.1002/wps.20472] [Medline: 28941113]
18. Firth J, Torous J, Nicholas J, Carney R, Rosenbaum S, Sarris J. Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. J Affect Disord 2017;218:15-22 [FREE Full text] [doi: 10.1016/j.jad.2017.04.046] [Medline: 28456072]
19. Wasil AR, Gillespie S, Patel R, Petre A, Venturo-Conerly KE, Shingleton RM, et al. Reassessing evidence-based content in popular smartphone apps for depression and anxiety: developing and applying user-adjusted analyses. J Consult Clin Psychol 2020;88(11):983-993. [doi: 10.1037/ccp0000604] [Medline: 32881542]
20. Beard C, Silverman AL, Forgeard M, Wilmer MT, Torous J, Björgvinsson T. Smartphone, social media, and mental health app use in an acute transdiagnostic psychiatric sample. JMIR Mhealth Uhealth 2019;7(6):e13364 [FREE Full text] [doi: 10.2196/13364] [Medline: 31199338]
21. Widnall E, Grant CE, Wang T, Cross L, Velupillai S, Roberts A, et al. User perspectives of mood-monitoring apps available to young people: qualitative content analysis. JMIR Mhealth Uhealth 2020;8(10):e18140 [FREE Full text] [doi: 10.2196/18140] [Medline: 33037875]
22. Torous J, Linardon J, Goldberg SB, Sun S, Bell I, Nicholas J, et al. The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. World Psychiatry 2025;24(2):156-174 [FREE Full text] [doi: 10.1002/wps.21299] [Medline: 40371757]
23. Lipschitz J, Miller CJ, Hogan TP, Burdick KE, Lippin-Foster R, Simon SR, et al. Adoption of mobile apps for depression and anxiety: cross-sectional survey study on patient interest and barriers to engagement. JMIR Ment Health 2019;6(1):e11334. [doi: 10.2196/11334] [Medline: 30681968]
24. Gindidis S, Stewart S, Roodenburg J. A systematic scoping review of adolescent mental health treatment using mobile apps. Advances in Mental Health 2018;17(2):161-177. [doi: 10.1080/18387357.2018.1523680]
25. Baumeister H, Reichler L, Munzinger M, Lin J. The impact of guidance on Internet-based mental health interventions — a systematic review. Internet Interventions 2014;1(4):205-215. [doi: 10.1016/j.invent.2014.08.003]
26. Duarte-Díaz A, Perestelo-Pérez L, Gelabert E, Robles N, Pérez-Navarro A, Vidal-Alaball J, et al. Efficacy, safety, and evaluation criteria of mhealth interventions for depression: systematic review. JMIR Ment Health 2023;10:e46877 [FREE Full text] [doi: 10.2196/46877] [Medline: 37756042]
27. Garrido S, Millington C, Cheers D, Boydell K, Schubert E, Meade T, et al. What works and what doesn't work? A systematic review of digital mental health interventions for depression and anxiety in young people. Front Psychiatry 2019;10:759 [FREE Full text] [doi: 10.3389/fpsyt.2019.00759] [Medline: 31798468]
28. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019;18(3):325-336 [FREE Full text] [doi: 10.1002/wps.20673] [Medline: 31496095]

29. Weisel KK, Fuhrmann LM, Berking M, Baumeister H, Cuijpers P, Ebert DD. Standalone smartphone apps for mental health-a systematic review and meta-analysis. NPJ Digit Med 2019;2:118 [FREE Full text] [doi: 10.1038/s41746-019-0188-8] [Medline: 31815193]
30. Martin DJ, Garske JP, Davis MK. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. J Consult Clin Psychol 2000;68(3):438-450. [doi: 10.1037//0022-006x.68.3.438]
31. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]
32. Calvert E, Cipriani M, Chen K, Dhima A, Burns J, Torous J. Evaluating clinical outcomes for anxiety and depression: a real-world comparison of the digital clinic and primary care. J Affect Disord 2025;377:275-283. [doi: 10.1016/j.jad.2025.02.051] [Medline: 39988138]
33. Macrynikola N, Nguyen N, Lane E, Yen S, Torous J. The digital clinic: an innovative mental health care delivery model utilizing hybrid synchronous and asynchronous treatment. NEJM Catalyst 2023;4(9). [doi: 10.1056/cat.23.0100]
34. Youn SJ, Schuler K, Sah P, Jaso-Yim B, Pennine M, O'Dea H, et al. Scaling out a digital-first behavioral health care model to primary care. Adm Policy Ment Health 2025;52(6):1036-1056. [doi: 10.1007/s10488-025-01433-2] [Medline: 40019640]
35. Patient Apps for Improved Healthcare: From Novelty to Mainstream.: IMS Institute for Healthcare Informatics; 2013. URL: https://ignacioriesgo.es/wp-content/uploads/2014/03/iihi_patient_apps_report_editora_39_2_1.pdf [accessed 2025-03-01]
36. Myers A, Chesebrough L, Hu R, Turchioe M, Pathak J, Creber R. Evaluating commercially available mobile apps for depression self-management. AMIA Annu Symp Proc 2020;2020:906-914 [FREE Full text] [Medline: 33936466]
37. Rickard NS, Kurt P, Meade T. Systematic assessment of the quality and integrity of popular mental health smartphone apps using the American Psychiatric Association's app evaluation model. Front Digit Health 2022;4:1003181 [FREE Full text] [doi: 10.3389/fdgth.2022.1003181] [Medline: 36246848]
38. Morello K, Schäfer SK, Kunzler AM, Priesterroth L, Tüscher O, Kubiak T. Cognitive reappraisal in mHealth interventions to foster mental health in adults: a systematic review and meta-analysis. Front Digit Health 2023;5:1253390 [FREE Full text] [doi: 10.3389/fdgth.2023.1253390] [Medline: 37927578]
39. Bakker D, Rickard N. Engagement with a cognitive behavioural therapy mobile phone app predicts changes in mental health and wellbeing: MoodMission. Aust Psychol 2020;54(4):245-260. [doi: 10.1111/ap.12383]
40. Seegan PL, Miller MJ, Heliste JL, Fathi L, McGuire JF. Efficacy of stand-alone digital mental health applications for anxiety and depression: a meta-analysis of randomized controlled trials. J Psychiatr Res 2023;164:171-183. [doi: 10.1016/j.jpsychires.2023.06.019] [Medline: 37352813]
41. Leigh S, Ashall-Payne L. The role of health-care providers in mHealth adoption. Lancet Digit Health 2019;1(2):e58-e59. [doi: 10.1016/S2589-7500(19)30025-1] [Medline: 33323231]
42. Schueller SM, Washburn JJ, Price M. Exploring mental health providers' interest in using web and mobile-based tools in their practices. Internet Interv 2016;4(2):145-151 [FREE Full text] [doi: 10.1016/j.invent.2016.06.004] [Medline: 28090438]
43. Henson P, David G, Albright K, Torous J. Deriving a practical framework for the evaluation of health apps. Lancet Digit Health 2019;1(2):e52-e54 [FREE Full text] [doi: 10.1016/S2589-7500(19)30013-5] [Medline: 33323229]
44. Torous J, Hsin H. Empowering the digital therapeutic relationship: virtual clinics for digital health interventions. NPJ Digit Med 2018;1(1):16 [FREE Full text] [doi: 10.1038/s41746-018-0028-2] [Medline: 31304301]
45. Lee M, Bin Mahmood ABS, Lee ES, Smith HE, Tudor Car L. Smartphone and mobile app use among physicians in clinical practice: scoping review. JMIR Mhealth Uhealth 2023;11:e44765 [FREE Full text] [doi: 10.2196/44765] [Medline: 37000498]
46. Dicianno BE, Salh A, Morris L, Xiang Y, Ding D. Rehabilitation clinicians' use of mainstream wireless technologies in practice: a scoping review. Disabil Rehabil Assist Technol 2024;19(8):2742-2760. [doi: 10.1080/17483107.2024.2316891] [Medline: 38349177]
47. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018;18(1):143 [FREE Full text] [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
48. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
49. Open Science Framework. URL: https://doi.org/10.17605/OSF.IO/VNY7J [accessed 2025-12-26]
50. Cooke A, Smith D, Booth A. Beyond PICO: the SPIDER tool for qualitative evidence synthesis. Qual Health Res 2012;22(10):1435-1443. [doi: 10.1177/1049732312452938] [Medline: 22829486]
51. Adams Z, Grant M, Hupp S, Scott T, Feagans A, Phillips ML, et al. Acceptability of an mHealth app for youth with substance use and mental health needs: iterative, mixed methods design. JMIR Form Res 2021;5(12):e30268. [doi: 10.2196/30268] [Medline: 34951593]
52. Almadani AH, Aldawood BD, Alahmari FM, AbuDujain NM, Otayf MM. Use and perceptions of mobile mental health applications among healthcare workers in Saudi Arabia: a cross-sectional study. J Nerv Ment Dis 2025;213(1):7-21. [doi: 10.1097/NMD.0000000000001812] [Medline: 39607414]

53. Anastasiadou D, Folkvord F, Serrano-Troncoso E, Lupiañez-Villanueva F. Mobile health adoption in mental health: user experience of a mobile health app for patients with an eating disorder. JMIR Mhealth Uhealth 2019;7(6):e12920 [FREE Full text] [doi: 10.2196/12920] [Medline: 31199329]
54. Armstrong CC, Odukoya EJ, Sundaramurthy K, Darrow SM. Youth and provider perspectives on behavior-tracking mobile apps: qualitative analysis. JMIR Ment Health 2021;8(4):e24482. [doi: 10.2196/24482] [Medline: 33885364]
55. Bucci S, Berry N, Morris R, Berry K, Haddock G, Lewis S, et al. "They Are Not Hard-to-Reach Clients. We Have Just Got Hard-to-Reach Services." staff views of digital health tools in specialist mental health services. Front Psychiatry 2019;10:344 [FREE Full text] [doi: 10.3389/fpsyt.2019.00344] [Medline: 31133906]
56. Chang S, Gray L, Alon N, Torous J. Patient and clinician experiences with sharing data visualizations integrated into mental health treatment. Social Sciences 2023;12(12):648 [FREE Full text] [doi: 10.3390/socsci12120648]
57. Cheung BS, Murphy JK, Michalak EE, Liu J, Yang X, Wang X, et al. Barriers and facilitators to technology-enhanced measurement based care for depression among Canadian clinicians and patients: results of an online survey. J Affect Disord 2023;320:1-6. [doi: 10.1016/j.jad.2022.09.055] [Medline: 36162664]
58. Deady M, Collins D, Gayed A, Harvey SB, Bryant R. The development of a smartphone app to enhance post-traumatic stress disorder treatment in high-risk workers. Digit Health 2023;9:20552076231155680. [doi: 10.1177/20552076231155680] [Medline: 36845080]
59. Dobson R, Variava R, Douglas M, Reynolds LM. Digital competency of psychologists in Aotearoa New Zealand: a cross-sectional survey. Front Digit Health 2022;4:951366 [FREE Full text] [doi: 10.3389/fdgth.2022.951366] [Medline: 36158995]
60. Dominiak M, Gędek A, Antosik AZ, Mierzejewski P. Mobile health for mental health support: a survey of attitudes and concerns among mental health professionals in Poland over the period 2020-2023. Front Psychiatry 2024;15:1303878. [doi: 10.3389/fpsyt.2024.1303878] [Medline: 38559395]
61. Dubad M, Elahi F, Marwaha S. The clinical impacts of mobile mood-monitoring in young people with mental health problems: the memo study. Front Psychiatry 2021;12:687270 [FREE Full text] [doi: 10.3389/fpsyt.2021.687270] [Medline: 34393850]
62. Etingen B, Zocchi MS, Higashi RT, Palmer JA, Richardson E, Bixler FR, et al. Mental health provider perspectives on a mobile health application to support remote measurement-based care: challenges and impacts. Psychol Serv 2025;22(2):243-255. [doi: 10.1037/ser0000884] [Medline: 39172403]
63. Francese R, Attanasio P. Emotion detection for supporting depression screening. Multimed Tools Appl 2023;82(9):12771-12795 [FREE Full text] [doi: 10.1007/s11042-022-14290-0] [Medline: 36570729]
64. González-Pérez A, Diaz-Sanahuja L, Matey-Sanz M, Osma J, Granell C, Bretón-López J, et al. Towards a self-applied, mobile-based geolocated exposure therapy software for anxiety disorders: SyMptOMS-ET app. Digit Health 2024;10:20552076241283942 [FREE Full text] [doi: 10.1177/20552076241283942] [Medline: 39484648]
65. Green JB, Rodriguez J, Keshavan M, Lizano P, Torous J. Implementing technologies to enhance coordinated specialty care framework: implementation outcomes from a development and usability study. JMIR Form Res 2023;7:e46491 [FREE Full text] [doi: 10.2196/46491] [Medline: 37788066]
66. Heydarian S, Shakiba A, Rostam Niakan Kalhori S. The minimum feature set for designing mobile apps to support bipolar disorder-affected patients: proposal of essential functions and requirements. J Healthc Inform Res 2023;7(2):254-276 [FREE Full text] [doi: 10.1007/s41666-023-00134-5] [Medline: 37377634]
67. Hildebrand AS, Planert J, Machulska A, Margraf LM, Roesmann K, Klucken T. Exploring psychotherapists' attitudes on internet- and mobile-based interventions in Germany: thematic analysis. JMIR Form Res 2024;8:e51832 [FREE Full text] [doi: 10.2196/51832] [Medline: 39510514]
68. Hoffman L, Benedetto E, Huang H, Grossman E, Kaluma D, Mann Z, et al. Augmenting mental health in primary care: A 1-year study of deploying smartphone apps in a multi-site primary care/behavioral health integration program. Front Psychiatry 2019;10:94 [FREE Full text] [doi: 10.3389/fpsyt.2019.00094] [Medline: 30873053]
69. Kerst A, Zielasek J, Gaebel W. Smartphone applications for depression: a systematic literature review and a survey of health care professionals' attitudes towards their use in clinical practice. Eur Arch Psychiatry Clin Neurosci 2020;270(2):139-152. [doi: 10.1007/s00406-018-0974-3] [Medline: 30607530]
70. Khan W, Jebanesan B, Ahmed S, Trimmer C, Agic B, Safa F, et al. Stakeholders' views and opinions on existing guidelines on "How to Choose Mental Health Apps". Front Public Health 2023;11:1251050 [FREE Full text] [doi: 10.3389/fpubh.2023.1251050] [Medline: 38074730]
71. Li S, Achilles M, Spanos S, Habak S, Werner-Seidler A, O'Dea B. A cognitive behavioural therapy smartphone app for adolescent depression and anxiety: co-design of ClearlyMe. Cogn Behav Therapist 2022;15 [FREE Full text] [doi: 10.1017/s1754470x22000095]
72. Lukka L, Karhulahti VM, Palva JM. Factors affecting digital tool use in client interaction according to mental health professionals: interview study. JMIR Hum Factors 2023;10:e44681 [FREE Full text] [doi: 10.2196/44681] [Medline: 37428520]

73. McGee-Vincent P, Mackintosh MA, Jamison AL, Juhasz K, Becket-Davenport C, Bosch J, et al. Training staff across the veterans affairs health care system to use mobile mental health apps: a national quality improvement project. JMIR Ment Health 2023;10:e41773 [FREE Full text] [doi: 10.2196/41773] [Medline: 36633895]
74. Medich M, Cannedy SL, Hoffmann LC, Chinchilla MY, Pila JM, Chassman SA, et al. Clinician and patient perspectives on the use of passive mobile monitoring and self-tracking for patients with serious mental illness: user-centered approach. JMIR Hum Factors 2023;10:e46909 [FREE Full text] [doi: 10.2196/46909] [Medline: 37874639]
75. Miller K, Kuhn E, Yu J, Owen J, Jaworski B, Taylor K, et al. Use and perceptions of mobile apps for patients among VA primary care mental and behavioral health providers. Prof Psychol Res Pract 2019;50(3):204-209 [FREE Full text] [doi: 10.1037/pro0000229]
76. Morton E, Torous J, Murray G, Michalak EE. Using apps for bipolar disorder - an online survey of healthcare provider perspectives and practices. J Psychiatr Res 2021;137:22-28. [doi: 10.1016/j.jpsychires.2021.02.047] [Medline: 33647725]
77. Naccache B, Mesquida, Raynaud JP, Revet A. Smartphone application for adolescents with anorexia nervosa: an initial acceptability and user experience evaluation. BMC Psychiatry 2021;21(1):467 [FREE Full text] [doi: 10.1186/s12888-021-03478-7] [Medline: 34563166]
78. Nogueira-Leite D, Diniz JM, Cruz-Correia R. Mental Health professionals' attitudes toward digital mental health apps and implications for adoption in portugal: mixed methods study. JMIR Hum Factors 2023;10:e45949. [doi: 10.2196/45949] [Medline: 37266977]
79. Orengo-Aguayo RE, Hanson RF, Moreland AD, Jobe-Shields L, Adams ZW. Enhancing the delivery of an empirically-supported trauma-focused treatment for adolescents: providers' views of the role of technology and web-based resources. Adm Policy Ment Health 2018;45(4):575-586. [doi: 10.1007/s10488-017-0846-6] [Medline: 29305776]
80. Patoz MC, Hidalgo-Mazzei D, Blanc O, Verdolini N, Pacchiarotti I, Murru A, et al. Patient and physician perspectives of a smartphone application for depression: a qualitative study. BMC Psychiatry 2021;21(1):65 [FREE Full text] [doi: 10.1186/s12888-021-03064-x] [Medline: 33514333]
81. Puhy C, Litke S, Silverstein M, Kiely J, Pardes A, McGeoch E, et al. Counselor and student perceptions of an mHealth technology platform used in a school counseling setting. Psychol Sch 2021;58(7):1284-1298 [FREE Full text] [doi: 10.1002/pits.22541]
82. Richards P, Simpson S, Bastiampillai T, Pietrabissa G, Castelnuovo G. The impact of technology on therapeutic alliance and engagement in psychotherapy: the therapist's perspective. Clinical Psychologist 2020;22(2):171-181 [FREE Full text] [doi: 10.1111/cp.12102]
83. Rodriguez-Villa E, Rozatkar AR, Kumar M, Patel V, Bondre A, Naik SS, et al. Cross cultural and global uses of a digital mental health app: results of focus groups with clinicians, patients and family members in India and the United States. Glob Ment Health (Camb) 2021;8:e30 [FREE Full text] [doi: 10.1017/gmh.2021.28] [Medline: 34512999]
84. Rothmann MJ, Mouritsen JD, Ladefoged NS, Jeppesen MN, Lillevang AS, Laustrup H, et al. The use of telehealth for psychological counselling of vulnerable adult patients with rheumatic diseases or diabetes: explorative study inspired by participatory design. JMIR Hum Factors 2022;9(1):e30829 [FREE Full text] [doi: 10.2196/30829] [Medline: 35311690]
85. Stefancic A, Rogers RT, Styke S, Xu X, Buchsbaum R, Nossel I, et al. Development of the first episode digital monitoring mhealth intervention for people with early psychosis: qualitative interview study with clinicians. JMIR Ment Health 2022;9(11):e41482 [FREE Full text] [doi: 10.2196/41482] [Medline: 36331539]
86. Strodl E, Shakespeare-Finch J, Alichniewicz KK, Brown K, Quinn C, Hides L, et al. Clinicians' perceptions of PTSD Ccoach Australia. Internet Interv 2020;21:100333. [doi: 10.1016/j.invent.2020.100333] [Medline: 32939341]
87. Weermeijer JDM, Wampers M, de Thurah L, Bonnier R, Piot M, Kuppens P, et al. Usability of the experience sampling method in specialized mental health care: pilot evaluation study. JMIR Form Res 2023;7:e48821 [FREE Full text] [doi: 10.2196/48821] [Medline: 37988137]
88. Wu DTY, Xin C, Bindhu S, Xu C, Sachdeva J, Brown JL, et al. Clinician perspectives and design implications in using patient-generated health data to improve mental health practices: mixed methods study. JMIR Form Res 2020;4(8):e18123 [FREE Full text] [doi: 10.2196/18123] [Medline: 32763884]
89. National Digital Mental Health Framework.: Australian Government Department of Health; 2021. URL: https://www.health.gov.au/sites/default/files/documents/2022/03/national-digital-mental-health-framework_0.pdf [accessed 2025-12-23]
90. Psychologists Struggle to Meet Demand Amid Mental Health Crisis: 2022 COVID-19 Practitioner Impact Survey.: American Psychological Association; 2022. URL: https://www.apa.org/pubs/reports/practitioner/2022-covid-psychologist-workload.pdf [accessed 2025-12-23]
91. Flores M, Glusman G, Brogaard K, Price ND, Hood L. P4 medicine: how systems medicine will transform the healthcare sector and society. Per Med 2013;10(6):565-576 [FREE Full text] [doi: 10.2217/pme.13.57] [Medline: 25342952]
92. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
93. Ng MM, Firth J, Minen M, Torous J. User engagement in mental health apps: a review of measurement, reporting, and validity. Psychiatr Serv 2019;70(7):538-544 [FREE Full text] [doi: 10.1176/appi.ps.201800519] [Medline: 30914003]

94. Luxton DD, McCann RA, Bush NE, Mishkind MC, Reger GM. mHealth for mental health: integrating smartphone technology in behavioral healthcare. Prof Psychol Res Pract 2011;42(6):505-512. [doi: 10.1037/a0024485]
95. Wang K, Varma DS, Prosperi M. A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders. J Psychiatr Res 2018;107:73-78. [doi: 10.1016/j.jpsychires.2018.10.006] [Medline: 30347316]
96. Entzeridou E, Markopoulou E, Mollaki V. Public and physician's expectations and ethical concerns about electronic health record: benefits outweigh risks except for information security. Int J Med Inform 2018;110:98-107. [doi: 10.1016/j.ijmedinf.2017.12.004] [Medline: 29331259]
97. Pote H, Rees A, Holloway-Biddle C, Griffith E. Workforce challenges in digital health implementation: how are clinical psychology training programmes developing digital competences? Digit Health 2021;7:2055207620985396 [FREE Full text] [doi: 10.1177/2055207620985396] [Medline: 33628457]
98. Bear HA, Ayala Nunes L, DeJesus J, Liverpool S, Moltrecht B, Neelakantan L, et al. Determination of markers of successful implementation of mental health apps for young people: systematic review. J Med Internet Res 2022;24(11):e40347. [doi: 10.2196/40347] [Medline: 36350704]
99. Connolly SL, Hogan TP, Shimada SL, Miller CJ. Leveraging implementation science to understand factors influencing sustained use of mental health apps: a narrative review. J Technol Behav Sci 2021;6(2):184-196 [FREE Full text] [doi: 10.1007/s41347-020-00165-4] [Medline: 32923580]

Abbreviations

APA: American Psychiatric Association
IMS-11: IMS Institute for Healthcare Informatics functionality score
MHapps: mental health smartphone apps
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analysis extension for Scoping Reviews
SPIDER: Sample, Phenomenon of Interest, Design, Evaluation, Research Type

Edited by S Sun; submitted 08.Apr.2025; peer-reviewed by M Barwick, D Hidalgo-Mazzei; comments to author 26.Aug.2025; accepted 09.Nov.2025; published 15.Jan.2026.

Please cite as:
Rickard NS, Kurt P, Meade T
Navigating the Digital Landscape for Potential Use of Mental Health Apps in Clinical Practice: Scoping Review
JMIR Ment Health 2026;13:e75640
URL: https://mental.jmir.org/2026/1/e75640
doi:10.2196/75640
PMID:

Original Paper

Advancing Psychiatric Safety With the Predictive Risk Identification for Mental Health Events Tool: Retrospective Cohort Study

Elham Dolatabadi[1,2], BSc, MSc, PhD; Valentina Tamayo Velasquez[3,4], HBsc; Abdul Hamid Dabboussi[1], BSc, MSc; David Wen[4], PhD; Jennifer Crawford[4,5], PhD; Andrea E Waddell[4,6], MEd, MD; Christo El Morr[1,7], BA, HBsc, MSc, PhD

[1]School of Health Policy and Management, Faculty of Health, York University, Toronto, ON, Canada
[2]Vector Institute, Toronto, ON, Canada
[3]Institute of Medical Science, Temerty Faculty Medicine, University of Toronto, Toronto, ON, Canada
[4]Waypoint Centre for Mental Health Care, Penetanguishene, ON, Canada
[5]Faculty of Health Sciences, The University of Ontario Institute of Technology, Oshawa, ON, Canada
[6]Temerty Faculty of Medicine, The University of Toronto, Toronto, ON, Canada
[7]North York General Hospital, Toronto, ON, Canada

Corresponding Author:
Elham Dolatabadi, BSc, MSc, PhD
School of Health Policy and Management
Faculty of Health
York University
Stong College, 340
165 Campus Walk
Toronto, ON, M3J 1P3
Canada
Phone: 1 6477069756
Email: edolatab@yorku.ca

Abstract

Background: Patient safety incidents are a leading cause of harm in psychiatric settings, yet early warning systems (EWS) tailored to mental health remain underdeveloped. Traditional risk tools such as the Dynamic Appraisal of Situational Aggression–Inpatient Version (DASA-IV) offer limited predictive accuracy and are reactive rather than proactive.

Objective: We introduce the Predictive Risk Identification for Mental Health Events (PRIME) tool, a deep learning–based EWS trained on longitudinal psychiatric electronic medical record (EMR) data to anticipate adverse events in 24-hour windows.

Methods: A retrospective cohort study using routinely collected EMR data to train and validate machine learning (ML) models for short-term risk prediction was conducted. This study took place at Waypoint Centre for Mental Health Care, a large inpatient psychiatric hospital in Ontario, Canada, serving both high-security forensic and nonforensic patient populations. A total of 4651 patients and 403,098 encounters from January 2020 to August 2024 were included. For model evaluation, the 2024 test set included 900 patients and 48,313 encounters. PRIME was trained using recurrent neural networks with attention mechanisms on multivariate time-series data. The model used an autoregressive design to forecast risk based on 7 days of prior patient data and was benchmarked against the DASA-IV clinical tool and other ML baselines. The primary outcome was the occurrence of an adverse mental health event recorded in the EMR within the following 24 hours. Model performance was assessed using area under the receiver operating characteristic curve (AUC) and recall, alongside subgroup analyses and interpretability assessments using integrated gradients.

Results: The long short-term memory with attention mechanism achieved the highest predictive performance (AUC=0.83), outperforming existing tools such as DASA-IV by 0.20 AUC (0.81 vs 0.61) and demonstrating the potential of ML-based models to support proactive risk management in mental health settings.

Conclusions: The PRIME tool is one of the first developed and evaluated deep learning–based EWS for psychiatric inpatient care. By outperforming existing clinical tools and providing interpretable, rolling predictions, PRIME offers a pathway toward safer, more proactive mental health interventions. Future work should assess its equity implications and integration into routine psychiatric workflows.

(JMIR Ment Health 2026;13:e84318) doi:10.2196/84318

KEYWORDS

early warning system; psychiatric adverse events; adverse event prediction; machine learning; clinical decision support; patient safety

Introduction

Patient and staff safety are top priorities in health care, yet patient safety incidents remain the third leading cause of death in Canada [1]. Many of these incidents stem from adverse events such as falls, medication errors, and medical complications [2]. A recent study found that 1 in 4 hospital admissions involved adverse events, with a quarter of these deemed preventable [3]. While all health care settings face safety risks, psychiatric environments present a distinct set of challenges, including suicide, restraint, and seclusion—events that contribute to continued deterioration and injury [4,5]. Despite a higher prevalence of adverse events in mental health, research on patient safety and mental deterioration–related adverse events in these settings remains limited compared to other medical fields [6,7].

These incidents not only worsen patient outcomes but also increase risks for staff [8]. Worldwide, approximately 24% of health care workers experience physical violence annually, with psychiatric staff at particularly high risk [9-11]. Reducing adverse events through assessment and prediction is crucial for improving staff and patient safety.

Current methods for assessing patient deterioration rely heavily on voluntary reporting, critical incident reviews, and clinician judgment [12]. While actuarial tools such as the Dynamic Appraisal of Situational Aggression and the Brøset Violence Checklist are also used, these 2 primarily target short-term aggression and violence prediction, and they have shown limited predictive accuracy and tend to miss early warning signs [13,14]. As a result, many opportunities for timely intervention are lost, especially in high-risk but low-observable cases with early signs of deterioration that are not easily detected. Additionally, there are other widely validated measures for more specific feature prediction, such as the Historical, Clinical, and Risk Management, also used for violence risk assessment; the Columbia Suicide Severity Rating Scale for the assessment of suicidal ideation and behavior; and many other risk assessment tools [15,16]. We focused on both the Dynamic Appraisal of Situational Aggression and Brøset Violence Checklist measures as they are 2 of the most widely validated and routinely implemented structured risk assessment tools in inpatient psychiatry [17].

Early warning systems (EWSs) are widely used in medicine, leveraging routinely collected clinical data to detect early signs of patient deterioration. Tools such as the National Early Warning Score 2 have been effectively implemented in acute care settings to support timely interventions [18-20]. At the same time, machine learning (ML) is transforming risk assessment by enabling the analysis of large-scale, high-dimensional health care data [21-23]. Predictive ML models are developed using historical patient records combined with expert input to train, test, and refine algorithms for higher performance and clinical relevance [24,25]. Compelling examples include CHARTWatch, developed to predict inpatient deterioration in general internal medicine, and Sepsis Watch, designed to identify patients at risk of sepsis before clinical recognition [26-28]. However, psychiatric care has not seen comparable innovation, in part due to the complexity of mental health data, lack of validated digital tools, and underrepresentation of psychiatric settings in EWS research.

To address this gap, we introduce a novel ML-based EWS, the Predictive Risk Identification for Mental Health Events (PRIME) tool. The PRIME tool is a deep learning–based EWS leveraging longitudinal electronic medical record (EMR) data from a specialized psychiatric hospital. The goal of the PRIME tool is to predict mental health–specific adverse events, including but not limited to self-harm, suicide attempts, violence toward others, and aggressive behaviors (Multimedia Appendix 1). PRIME is trained to predict the likelihood of these adverse events within 24-hour windows using autoregressive recurrent neural networks enriched with attention mechanisms and interpretability via integrated gradients. Unlike traditional tools, PRIME is capable of continuous, real-time risk forecasting even in the absence of prior incidents. We benchmarked PRIME against Dynamic Appraisal of Situational Aggression–Inpatient Version (DASA-IV) and other ML models, and it demonstrated superior performance, particularly in complex and high-risk subgroups.

Through this study, we aimed to move beyond reactive safety practices toward proactive, data-informed risk mitigation in mental health care, advancing both patient and staff safety in a setting long underserved by digital innovation.

Methods

Study Design and Data Acquisition

In this study, we used routinely collected clinical data extracted from the EMR at Waypoint Centre for Mental Health Care (hereafter referred to as “Waypoint”), Ontario, Canada. We retrospectively retrieved data from all patients at Waypoint between January 2020 and August 2024, including static and dynamic variables (Multimedia Appendix 1).

Ethical Considerations

This study was approved by the York University Office of Research Ethics (certificate e2023-163) and the Research Ethics Board of Waypoint Centre for Mental Health Care (reference #RCRA#23.08.01) with waived informed consent. The Research Ethics Board waived the need for informed consent since the data was retrospectively collected in routine practice.

Data Representation and Processing

First, we conducted a literature review to identify factors widely associated with mental health deterioration and adverse events. We collaborated with clinicians, physicians, clinical informatics specialists, and the research team to review these factors and select the variables within our EMR (Multimedia Appendix 1). The baseline data preprocessing included one-hot encoding and normalization of all measures. We implemented a standardized aggregation strategy to address the variability in time-series data arising from differing measurement frequencies, where some clinical parameters were recorded daily and others were recorded multiple times per day. These factors encompassed a range of clinical and behavioral variables in the following categories: inpatient admission assessments that included demographic and diagnosis data, clinical risk assessments, physiological data, recent behavioral data, and mental status assessment data (Multimedia Appendix 1). Patient encounters were segmented into 24-hour intervals, aligning with clinical workflows that typically operate on daily cycles for alerts. Within each interval, all measures were aggregated to provide a comprehensive snapshot of patient health over the specified time frame. Numerical variables were averaged across the interval, whereas categorical variables were first encoded numerically based on severity or clinical importance and then summed within the 24-hour period.

We collected admission diagnosis data based on the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*, selecting the 45 most frequent diagnoses across all patients to prevent overfitting. Medication data included our patient group's 5 most relevant categories, each represented as a binary indicator denoting whether it was administered within the previous 24 hours (Multimedia Appendix 1).

The primary outcome in our study was the occurrence of any mental health adverse event. For our prediction task, each patient encounter was labeled based on whether a logged adverse event in the EMR system occurred within the following 24-hour bin. This binary label (event vs no event) was used as the target variable for PRIME training. Moreover, prior adverse events in the previous 24-hour intervals were also dynamically added to future intervals, referred to as the history of any incident.

Building the PRIME Model

To improve psychiatric-medical baselines, we designed a deep learning–based EWS. Specifically, we developed recurrent neural networks using the long short-term memory (LSTM) model that triggered an alert every 24 hours based on a variable sequence length (3-7 days) of patient data, treated as a hyperparameter. During training, the ground truth history of each adverse event was provided for every 24-hour interval in the model. In the inference phase, the model operated in an autoregressive mode: it used its own predicted output for the previous 24-hour window as an input signal for the next prediction step. To enhance the model's ability to focus on the most relevant temporal signals within the input sequence, we further explored an LSTM model with attention mechanisms (LSTM+attention). In this variant, an attention layer was added to the LSTM hidden states Multimedia Appendix 2. For each 24-hour prediction interval, the attention mechanism dynamically assigned weights to each time step in the historical input sequence, allowing the model to selectively focus on the most informative data that contributed to the risk signal.

We also evaluated our model against 2 ML approaches: light gradient boosting machine (LightGBM) and feedforward neural network (FNN). All time-series features were aggregated over 3 to 7 days using the same methodology applied to 24-hour intervals. Predictions were then made for the next 24-hour interval, allowing for consistent model evaluation and direct comparison across different sequence lengths to identify the best-performing approach. To support robust evaluation and model selection, the dataset was first partitioned into two distinct test sets: (1) a held-out patient test set with no patient overlap between the development (3000 patients) and the test sets (751 patients) and (2) an out-of-time test set split across time using 2020 to 2022 data for training and 2023 data for testing.

Once the model selection was finalized, we performed a final training phase using all data from 2020 to 2023 to build PRIME. This final model was then evaluated on 2024 data to assess real-world applicability. Model calibration was assessed on the 2024 evaluation cohort using reliability (calibration) curves and the Brier score [29,30]. Reliability curves were generated using 10 uniformly spaced probability bins plotting the mean predicted risk against observed outcome frequency within each bin. The Brier score was computed as the mean squared difference between predicted probabilities and observed binary outcomes, providing a quantitative measure of overall probabilistic accuracy.

To explore the factors driving PRIME's predictions, we used integrated gradients to compute feature importance in our LSTM model by computing the path-integrated gradients from the input to the actual output [31]. To quantify uncertainty, gradients were bootstrapped over 100 resampled datasets.

Comparison With Clinical Measures

We compared PRIME's predictive performance against that of the DASA-IV, a standardized tool used at our hospital to evaluate risks of aggression [32]. The PRIME tool's predictions included all mental health–specific adverse events recorded in the hospital's incident log (Multimedia Appendix 1). DASA-IV includes 7 items assessing behavioral indicators (ie, irritability, negative attitudes, and verbal threats), each scored as 0 (not observed) or 1 (observed), with a total score categorized as low (0-1), moderate (2-3), or high (>3) [33,34]. To align DASA-IV with PRIME's binary classification, we restructured the risk categories. Moderate and high risk were grouped as "at risk" (positive prediction), whereas low risk was grouped as "no risk" (negative prediction). PRIME is designed to predict a broader range of mental health–specific adverse events, whereas DASA-IV is limited to aggression-related incidents and deterioration. Our goal was to compare the PRIME tool with the current validated tool used in clinical practice. This allowed us to compare DASA-IV's performance against PRIME's predictions and the ground truth outcomes recorded in patient encounters.

Results

Cohort Characteristics

The dataset encompassed 4651 patients and 403,098 patient encounters over 55 months. The demographic characteristic distribution of the patient cohort is presented in Table 1. For the evaluation of the best-performing ML model and comparison against clinical baselines, we used data from 2024, with detailed breakdowns provided in Table 1.

Table 1. Cohort characteristics and dataset splits used for model development and evaluation. After final model selection, the full dataset from 2020 to 2023 (model development) was used to train the Predictive Risk Identification for Mental Health Events, which was evaluated using 2024 (model evaluation) data to assess real-world performance.

	Model development (2020-2023)				Model evaluation (2024)—evaluation set (48,313)
	Held-out patients		Out-of-time patients		
Data split (number of patient encounters)	Development set (281,022)	Test set (73,763)	Development set (259,257)	Test set (95,528)	
Patients, n (%)	3000 (80)	751 (20)	2851 (70.3)	1202 (29.7)	900 (100)
Period	January 1, 2020, to December 31, 2023	January 1, 2020, to December 31, 2023	January 1, 2020, to December 31, 2022	January 1, 2023, to December 31, 2023	January 1, 2024, to August 19, 2024
LOS[a], mean (SD)	629.22 (632.67)	530.41 (654.51)	708.04 (676.56)	336.90 (390.40)	134.99 (117.83)
Sex, n (%)					
Female	839 (27.98)	231 (30.81)	845 (29.63)	310 (25.75)	235 (26.05)
Male	2045 (68.17)	502 (66.82)	1912 (67.08)	842 (70.05)	633 (70.37)
Other	116 (3.85)	18 (2.37)	94 (3.29)	50 (4.20)	32 (3.58)
Sexual orientation, n (%)					
Heterosexual	1878 (62.59)	472 (62.90)	1833 (64.31)	699 (58.15)	541 (60.14)
Other	1122 (37.41)	279 (37.10)	1018 (35.69)	503 (41.85)	359 (39.86)
Race, n (%)					
Black	273 (9.10)	36 (4.73)	224 (7.85)	109 (9.03)	73 (8.10)
First Nations	61 (2.05)	18 (2.45)	63 (2.21)	23 (1.92)	19 (2.13)
White	1987 (66.23)	572 (76.20)	2000 (70.15)	763 (63.49)	570 (63.28)
Other races	679 (22.62)	125 (16.62)	564 (19.78)	307 (25.57)	238 (26.49)
Incident prevalence					
Total number of incidents	11,744	2569	10,688	3625	2106
Patients, n (%)	762	209	766	342	266

[a]LOS: length of stay in days.

Adverse event distribution per individual varied across the sample of patients between 2020 to 2024. When grouping the number of patients by the frequency of adverse events they experienced during their hospital stay, there is a decrease in the number of patients who experience a high count of adverse events. Most patients experienced few or no adverse events: 69.9% (3251/4651) had no incidents, and 12.6% (587/4651) experienced up to 2 incidents. A total of 7.4% (344/4651) of the patients had between 3 and 16 events, with a median of 14.5 (IQR 14.25). A smaller group of 162 patients experienced between 17 and 83 incidents, most of whom (n=42, 25.9%) had between 17 and 20, whereas only 22 (13.6%) had more than 83 events. As incident frequency increased, cohort size decreased. The mean number of adverse events across the sample was 2.85, whereas the mode and median were both 0, highlighting the skewed nature of the data. This imbalance is important to consider as it affects how the model learns from the hospital's patient population, with most of the training data representing patients with few or no incidents.

PRIME's Predictions

Table 2 presents the performance comparison of the 4 ML models: light gradient boosting (LightGBM), feedforward neural network (FNN), LSTM, and LSTM+attention. Each model was trained multiple times using different random seeds, and performance metrics were averaged across runs to ensure robustness. Given the imbalanced nature of the dataset, model performance was evaluated using the area under the receiver operating characteristic curve (AUC) and recall. The LSTM+attention model consistently achieved the highest performance, with an AUC of 0.87 for held-out patients and 0.72 for out-of-time patients Multimedia Appendix 3. We selected the LSTM+attention model as the final architecture for PRIME (Table 2).

Table 2. Performance comparison of 4 machine learning models evaluated using area under the receiver operating characteristic curve (AUC) and recall. Metrics were averaged across multiple runs with different random seeds to ensure robustness.

Category and subcategory	AUC	Recall
Model selection, mean (SD)		
Held-out patients		
LightGBM[a]	0.51 (0.004)	0.02 (0.009)
FNN[b]	0.52 (0.005)	0.05 (0.011)
LSTM[c]	0.87 (0.002)	0.75 (0.02)
LSTM+attention	0.87 (0.002)	0.74 (0.04)
Out-of-time patients		
LightGBM	0.52 (0.002)	0.04 (0.003)
FNN	0.54 (0.01)	0.08 (0.03)
LSTM	0.84 (0.01)	0.72 (0.01)
LSTM+attention	0.85 (0.01)	0.75 (0.02)
PRIME's[d] performance		
Sex		
Male	0.83	0.36
Female	0.84	0.29
Intersex	0.87	0.23
Race		
Black	*0.69* [e]	0.16
First Nations	0.8	0.16
White	0.84	0.38
Other racial identities	*0.81*	0.27
Sexual orientation		
Heterosexual	0.82	0.34
Other	0.84	0.34
Program type		
Regional (nonforensic)	0.83	0.34
Provincial (forensic)	0.8	0.27
Age group (years)		
18-65	0.81	0.32
≥65	0.81	0.38
All	0.81	0.3

[a]LightGBM: light gradient boosting machine.

[b]FNN: feedforward neural network.

[c]LSTM: long short-term memory.

[d]PRIME: Predictive Risk Identification for Mental Health Events.

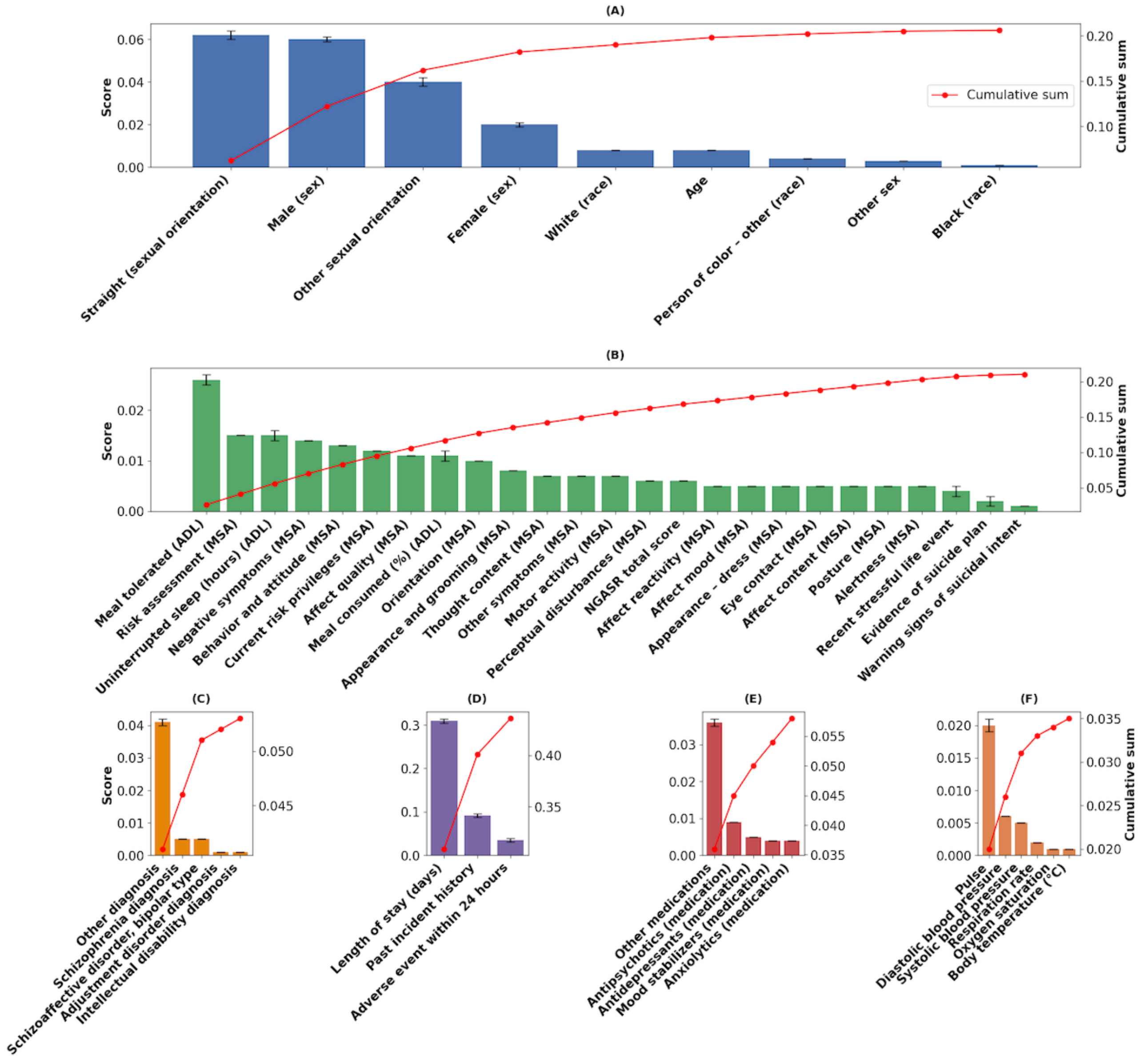
[e]Italicization indicates significance.

For the evaluation using the dataset from 2024, with 48,313 encounters and 2106 recorded adverse events, PRIME achieved an AUC of 83% (Table 2). The performance varied within and across subgroups, with AUC ranging from 0.69 (Black patients) to 0.87 (intersex patients), indicating potential biases favoring larger, more represented groups within the sample data. Across racial subgroups, AUC differed by 14%; across sex subgroups, AUC varied by 5%; across sexual orientation subgroups, AUC varied by 2%; across program types, AUC differed by 4%; and, across age groups, AUC variation was minimal (<1%).

Calibration analysis demonstrated that PRIME produced well-aligned risk estimates. The reliability curve closely followed the identity line across predicted probability bins (Multimedia Appendix 4), indicating good agreement between predicted and observed event rates. The model achieved a Brier score of 0.036 on the evaluation set, reflecting strong overall calibration performance given the low event prevalence.

Feature importance was aggregated across time steps and encounters and summarized at the feature level (Figure 1). Integrated gradient attributions were bootstrapped over 100 resampled evaluation datasets, with the resulting variability visualized as error bars. Ranking stability was assessed using Spearman correlation, demonstrating near-perfect robustness (ρ=0.99, –0.0001 to +0.0001). Among the 40 features included in PRIME, the top 16 predictors accounted for approximately 80% of the model's total importance, reflecting a diverse combination of demographic, medical, and psychosocial factors that drive risk prediction.

Figure 1. Feature importance across distinct categories in the PRIME model: (A) Demographic features including gender, race, and sexual orientation; (B) Clinical assessment features such as mental status indicators and functional assessments; (C) Clinical diagnoses including major psychiatric conditions, including schizophrenia diagnosis; (D) Clinical variables related to adverse events, incident history, and hospital stay duration; (E) Medication-related features, including mood stabilizers and antipsychotics; and (F) Vital signs, including pulse, blood pressure, and oxygen saturation.

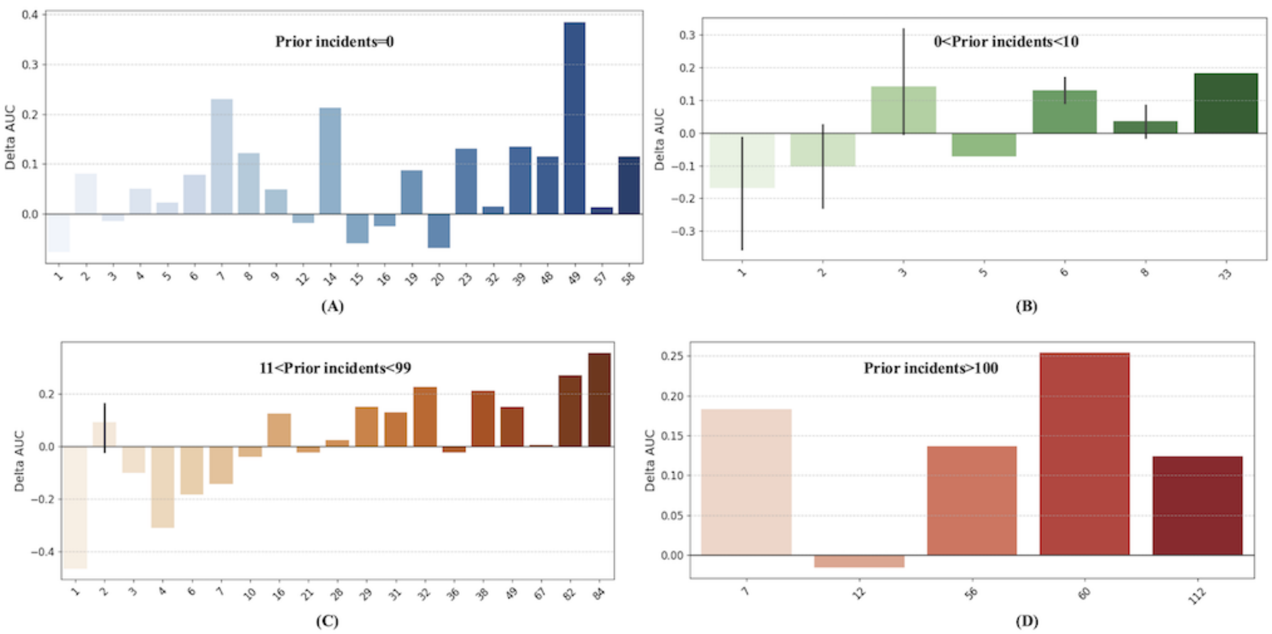


When further analyzing feature contributions within specific categories, demographic factors (Figure 1A) and indicators, with heterosexual sexual orientation and male sex showing the largest individual contributions, followed by *other* sexual orientation categories and female sex. Race-related variables and age demonstrated comparatively smaller effects. From the clinical assessments (Figure 1B), meal tolerated (ADL), risk assessment (MSA) and uninterrupted sleep (ADL) were identified as important contributors. In the category of clinical diagnoses (Figure 1C), schizophrenia and schizoaffective or bipolar disorder emerged as the most significant predictors. Among clinical variables (Figure 1D), length of hospital stay emerged as the most influential contributor, followed by history of past incidents and adverse events in the previous 24 hours. In the medication category (Figure 1E), the other medications category exhibited the largest overall contribution, followed by antipsychotics and antidepressants, with mood stabilizers and anxiolytics contributing more modestly. Finally, among vital signs (Figure 1F), features such as pulse, blood pressure, and oxygen saturation met the 0.90 cumulative importance threshold, although their influence remained relatively modest.

Comparison With the Standardized Risk Assessment Tool DASA-IV

PRIME demonstrated a 0.2 AUC improvement over the DASA-IV assessment tool when assessed on the 2024 evaluation dataset, with PRIME achieving an AUC of 0.81 compared to DASA-IV's AUC of 0.61 (Multimedia Appendix 5). To further assess PRIME's performance across different patient groups, we analyzed its effectiveness based on the historical incidence of adverse events for each individual in the training dataset (previous adverse event history). Figure 2 illustrates the performance differences between PRIME and DASA-IV across various patient groups, where each group is defined by the number of adverse events recorded in both the training (past) and evaluation (future) datasets. To examine the model's performance compared to that of DASA-IV, we defined subgroups based on all the unique combinations of adverse event occurrences observed in the training and test datasets. This yielded 63 unique subgroups representing different patterns and combinations of past and future incident frequencies across the datasets.

Figure 2. Difference in AUC ROC performance scores (Delta AUC = ML AUC ROC - DASA AUC ROC). A positive delta AUC indicates the ML model outperformed DASA for that specific cohort group. A negative delta AUC indicates DASA outperformed ML for that specific cohort group. AUC: area under the receiver operating characteristic curve; DASA: Dynamic Appraisal of Situational Aggression; ML: machine learning; ROC: receiver operating characteristic curve.



The PRIME tool significantly outperformed DASA-IV in 40 of the 63 subgroups (Wilcoxon test; *P*=.007). For individuals with no prior incidents in the training set but up to 58 total incidents in the evaluation period (Figure 2A), PRIME achieved an AUC of 0.62, whereas DASA-IV achieved an AUC of 0.50. For individuals with up to 10 incidents in the past and up to 23 in the future (Figure 2B), DASA-IV outperformed PRIME in cases in which individuals had 1, 2, or 5 future incidents. However, PRIME outperformed DASA-IV in the remaining 4 subcategories within this range. For individuals with moderate incident frequency (11-99 past incidents), PRIME outperformed DASA-IV in 11 of the 19 groups (Figure 2C). Among individuals with frequent incidents (>100 past incidents), PRIME outperformed DASA-IV in 4 of the 5 subgroups (Figure 2D). Notably, PRIME's performance was better in edge cases in which individuals had a high number of past incidents but only 1 in the future.

Discussion

Despite the growing number of adverse events in mental health settings, deep learning tools that leverage routinely collected EMR data to predict patient deterioration remain limited. Our model, PRIME, represents a first-of-its-kind approach tailored specifically to psychiatry and demonstrated strong predictive performance, achieving an AUC of 0.83. Leveraging autoregressive LSTM with attention mechanisms, PRIME operates in a rolling prediction mode, enabling 24-hour forecasts even in the absence of recent incident data. Notably, the history of prior incidents emerged as one of the most informative features, reinforcing the predictive value of temporal continuity in patient risk trajectories. Furthermore, the inclusion of patients from both forensic and nonforensic acute care programs contributes to the model's generalizability across diverse mental health populations. The strong calibration performance observed for PRIME is particularly important for clinical deployment, where accurate probability estimates are essential for risk stratification and decision support. Well-calibrated predictions enable clinicians to interpret PRIME scores as meaningful risk estimates rather than solely as ranking signals.

Currently, no ML-based predictive alerting tools are deployed in mental health settings. Instead, clinicians rely on actuarial tools such as DASA-IV to assess risks related to violence and aggression [35]. On the same dataset, PRIME outperformed DASA-IV (AUC=0.83 vs 0.61). While DASA-IV has reported AUCs between 0.61 and 0.82 in other studies, it is important to note that PRIME and DASA-IV target different outcomes [36]. PRIME captures a broader spectrum of deterioration events, including suicide, self-harm, and clinical decompensation, whereas DASA-IV is limited to aggression-related outcomes. The lower AUC for DASA-IV in our dataset likely reflects these differences in scope. Nonetheless, PRIME's ability to deliver significantly stronger performance across a wider range of adverse events underscores its versatility and robustness. In clinical practice, focusing solely on aggression is insufficient; risks of suicide and self-harm are equally critical. By encompassing a more comprehensive set of risks, PRIME provides clinicians with a holistic and actionable risk assessment framework, supporting earlier and more effective interventions. PRIME also showed strong performance even in patients with no prior recorded incidents, addressing a critical limitation of traditional tools that rely heavily on observable behavior or clinician judgment.

The feature "adverse event in the past 24 hours" emerged as one of the predictors of future deterioration, consistent with findings from acute care settings where recent clinical instability is a key driver of risk. Similar patterns have been observed in inpatient deterioration models, where temporal proximity to prior events significantly enhances predictive accuracy [37,38]. Beyond clinical history, our results indicate that a wide array of features, including demographic variables, mental status assessments, clinical diagnoses, medications, and vital signs, contribute meaningfully to risk prediction. This multidimensional pattern aligns with emerging work suggesting that accurate prediction of psychiatric outcomes requires integrating different types of structured medical data and psychosocial factors [39-41]. Overall, these findings underscore the importance of using holistic patient representations to capture the complex drivers of risk in mental health, a direction that has been underexplored in existing ML applications in psychiatry.

A limitation of this study, as previously noted, is the underrepresentation of certain demographic subgroups, which affected the model's predictive performance. We observed up to an 18% variation in AUC across subpopulations, indicating disparities in performance. Notably, the model was less accurate for 2 racial subgroups: Black and First Nations individuals, with AUC scores 14% and 3% lower, respectively, than those for the overall model performance. Additionally, both groups had a recall of 0.16, which was lower than that of all other subgroups, suggesting a higher rate of false negatives and an undercalling of risk. These disparities likely stem from the low representation of these groups in the dataset, with Black individuals comprising less than 10% (309/3751) of the sample in the training set and less than 10% (73/900) in the evaluation set. Similarly, First Nations individuals comprise less than 3% (79/3751) in the training set and less than 3% (19/900) in the evaluation set. Furthermore, this study did not assess the intersectional effects, such as whether the demographic factors had any effect or potential differences between the forensic and nonforensic programs. These represent important assessments for future work to further evaluate whether predictive models such as PRIME are unbiased and generalizable across different clinical settings and patient populations.

Additionally, while the PRIME tool demonstrated high predictive performance, the complexity that is inherently present in deep learning models may limit clinical interpretability. Ensuring clinician confidence and understanding of the model's prediction is critical for successful implementation. Ongoing monitoring and evaluation of PRIME are needed to assess its real-world performance and potential biases.

Future work will evaluate the utility, feasibility, and efficacy of the PRIME tool in real-world clinical settings. This future work will also focus on mitigating the previously mentioned biases through bias-aware data augmentation and fairness-aware learning algorithms (eg, adversarial debiasing) to improve representation across subgroups [42-44]. Piloting the PRIME tool in a live clinical setting is the next step in validating its performance and efficacy and informing the next steps toward broader clinical deployment. In our future pilot and deployment, we plan to use PRIME as a binary risk assessment tool to flag patients at a high risk of adverse events in mental health settings. Finally, although the PRIME tool was developed using data from a single mental health hospital, the model framework and variable-mapping methodological approaches are transferable to other mental health and psychiatric settings. If the PRIME tool is to be implemented in other settings, it will require retraining and validation to account for different patient populations, data sources, and documentation practices.

In this study, we developed and evaluated an LSTM model that could predict patients at risk of an adverse event. The model showed good performance across different subgroup populations, and our findings suggest that the model would outperform currently used risk assessment tools. Its autoregressive design, model evaluation, and near–real-time operation position it for real-world clinical integration. By generating dynamic forecasts without dependence on manual clinician input, PRIME can augment existing workflows and support earlier interventions in settings where mental health staff face high demands and elevated safety risks.

Acknowledgments

This work was made possible through access to data provided by Waypoint Centre for Mental Health Care, and the authors gratefully acknowledge the support and collaboration of the Waypoint research team. They also extend their appreciation to York University and its dynamic research community at the intersection of health, data science, and machine learning.

Funding

This work was funded by the Healthcare Insurance Reciprocal of Canada. ED's research is supported by the Canadian Institutes of Health Research Centre for Research on Pandemic Preparedness and Health Emergencies and a Natural Sciences and Engineering Research Council of Canada Discovery Grant.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to privacy policies and ethical restrictions but are available from the corresponding author on reasonable request.

Authors' Contributions

ED, AEW, and CEM contributed to the conceptualization and methodology of the study. ED, CEM, JC, AHD, and DW contributed to coding and model development. VTV and ED conducted the data analysis and writing of the manuscript. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of baseline and temporal features used for the PRIME (Predictive Risk Identification for Mental Health Events) tool.
[DOCX File , 4251 KB - mental_v13i1e84318_app1.docx]

Multimedia Appendix 2
Hyperparameters for the long short-term memory model and the long short-term memory+attention model. A comparison of the hyperparameter configurations for the two top performing models evaluated.
[DOCX File , 3742 KB - mental_v13i1e84318_app2.docx]

Multimedia Appendix 3
A summary of the model's performance metrics across 18 epochs. Early stopping triggered at epoch 19.
[DOCX File , 3743 KB - mental_v13i1e84318_app3.docx]

Multimedia Appendix 4
Calibration (reliability) curve for the PRIME model evaluated on the 2024 test dataset.
[DOCX File , 3917 KB - mental_v13i1e84318_app4.docx]

Multimedia Appendix 5
Clinical baseline model: Dynamic Appraisal of Situational Aggression's Predictive Performances across different data splits (area under the receiver operating characteristic curve).
[DOCX File , 3741 KB - mental_v13i1e84318_app5.docx]

References

1. The case for investing in patient safety in Canada. Canadian Patient Safety Institute. URL: https://www.bcit.ca/files/health/pdf/risk-analytica-2017-investing-in-patient-safety-in-canada.pdf [accessed 2025-05-29]
2. Leape LL, Brennan TA, Laird N, Lawthers AG, Localio AR, Barnes BA, et al. The nature of adverse events in hospitalized patients: results of the Harvard Medical Practice Study II. N Engl J Med 1991 Feb 07;324(6):377-384. [doi: 10.1056/nejm199102073240605]
3. Bates DW, Levine DM, Salmasian H, Syrowatka A, Shahian DM, Lipsitz S, et al. The safety of inpatient health care. N Engl J Med 2023 Jan 12;388(2):142-153. [doi: 10.1056/nejmsa2206117]
4. Hilton NZ, Ham E, Rodrigues NC, Kirsh B, Chapovalov O, Seto MC. Contribution of critical events and chronic stressors to PTSD symptoms among psychiatric workers. Psychiatr Serv 2020 Mar 01;71(3):221-227. [doi: 10.1176/appi.ps.201900226] [Medline: 31795856]

5. Chieze M, Hurst S, Kaiser S, Sentissi O. Effects of seclusion and restraint in adult psychiatry: a systematic review. Front Psychiatry 2019;10:491 [FREE Full text] [doi: 10.3389/fpsyt.2019.00491] [Medline: 31404294]
6. Waddell AE, Gratzer D. Patient safety and mental health-a growing quality gap in Canada. Can J Psychiatry 2022 Apr 11;67(4):246-249 [FREE Full text] [doi: 10.1177/07067437211036596] [Medline: 34378413]
7. Velasquez VT, Chang J, Waddell A. The development of early warning scores or alerting systems for the prediction of adverse events in psychiatric patients: a scoping review. BMC Psychiatry 2024 Oct 28;24(1):742 [FREE Full text] [doi: 10.1186/s12888-024-06052-z] [Medline: 39468486]
8. Lee JR, Kim EM, Kim SA, Oh EG. A systematic review of early warning systems' effects on nurses' clinical performance and adverse events among deteriorating ward patients. J Patient Saf 2020 Sep 26;16(3):e104-e113. [doi: 10.1097/PTS.0000000000000492] [Medline: 29698354]
9. Liu J, Gan Y, Jiang H, Li L, Dwyer R, Lu K, et al. Prevalence of workplace violence against healthcare workers: a systematic review and meta-analysis. Occup Environ Med 2019 Dec 13;76(12):927-937. [doi: 10.1136/oemed-2019-105849] [Medline: 31611310]
10. Hesketh KL, Duncan SM, Estabrooks CA, Reimer MA, Giovannetti P, Hyndman K, et al. Workplace violence in Alberta and British Columbia hospitals. Health Policy 2003 Mar;63(3):311-321. [doi: 10.1016/s0168-8510(02)00142-2] [Medline: 12595130]
11. Hiebert BJ, Care WD, Udod SA, Waddell CM. Psychiatric nurses' lived experiences of workplace violence in acute care psychiatric units in Western Canada. Issues Ment Health Nurs 2022 Feb 11;43(2):146-153. [doi: 10.1080/01612840.2021.1956656] [Medline: 34379570]
12. Hibbert PD, Molloy CJ, Schultz TJ, Carson-Stevens A, Braithwaite J. Comparing rates of adverse events detected in incident reporting and the Global Trigger Tool: a systematic review. Int J Qual Health Care 2023 Jul 25;35(3):mzad056 [FREE Full text] [doi: 10.1093/intqhc/mzad056] [Medline: 37440353]
13. Sammut D, Hallett N, Lees-Deutsch L, Dickens GL. A systematic review of violence risk assessment tools currently used in emergency care settings. J Emerg Nurs 2023 May;49(3):371-86.e5 [FREE Full text] [doi: 10.1016/j.jen.2022.11.006] [Medline: 36585335]
14. Ogonah MG, Seyedsalehi A, Whiting D, Fazel S. Violence risk assessment instruments in forensic psychiatric populations: a systematic review and meta-analysis. Lancet Psychiatry 2023 Oct;10(10):780-789 [FREE Full text] [doi: 10.1016/S2215-0366(23)00256-0] [Medline: 37739584]
15. Dolan M, Blattner R. The utility of the Historical Clinical Risk -20 scale as a predictor of outcomes in decisions to transfer patients from high to lower levels of security-a UK perspective. BMC Psychiatry 2010 Sep 29;10(1):76. [doi: 10.1186/1471-244x-10-76]
16. Xiao S, Ge Q, Wang T, Zhang M, Hu A, Zhang X. Psychometric characteristics of the Chinese version of the Columbia-Suicide Severity Rating Scale among people with mental health diagnosis. BMC Psychiatry 2025 Aug 21;25(1):803 [FREE Full text] [doi: 10.1186/s12888-025-07187-3] [Medline: 40842006]
17. Chu CM, Daffern M, Ogloff JR. Predicting aggression in acute inpatient psychiatric setting using BVC, DASA, and HCR-20 Clinical scale. J Forensic Psychiatry Psychol 2013 Apr;24(2):269-285. [doi: 10.1080/14789949.2013.773456]
18. Suspected Sepsis: Recognition, Diagnosis and Early Management. London, UK: National Institute for Health and Care Excellence; 2024.
19. Wibisono E, Hadi U, Arfijanto MV, Rusli M, Rahman BE, Asmarawati TP, et al. National early warning score (NEWS) 2 predicts hospital mortality from COVID-19 patients. Ann Med Surg (Lond) 2022 Apr;76:103462 [FREE Full text] [doi: 10.1016/j.amsu.2022.103462] [Medline: 35284070]
20. Zaidi H, Bader-El-Den M, McNicholas J. Using the National Early Warning Score (NEWS/NEWS 2) in different Intensive Care Units (ICUs) to predict the discharge location of patients. BMC Public Health 2019 Sep 05;19(1):1231 [FREE Full text] [doi: 10.1186/s12889-019-7541-3] [Medline: 31488143]
21. El Morr C, Jammal M, Ali-Hassan H, EI-Hallak W. Machine Learning for Practical Decision Making: A Multidisciplinary Perspective with Applications from Healthcare, Engineering and Business Analytics. Cham, Switzerland: Springer; 2022.
22. Menger V, Spruit M, van Est R, Nap E, Scheepers F. Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records. JAMA Netw Open 2019 Jul 03;2(7):e196709 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.6709] [Medline: 31268542]
23. Valli I, Marquand AF, Mechelli A, Raffin M, Allen P, Seal ML, et al. Identifying individuals at high risk of psychosis: predictive utility of support vector machine using structural and functional MRI data. Front Psychiatry 2016 Apr 08;7:52 [FREE Full text] [doi: 10.3389/fpsyt.2016.00052] [Medline: 27092086]
24. van der Vegt AH, Campbell V, Mitchell I, Malycha J, Simpson J, Flenady T, et al. Systematic review and longitudinal analysis of implementing artificial intelligence to predict clinical deterioration in adult hospitals: what is known and what remains uncertain. J Am Med Inform Assoc 2024 Jan 18;31(2):509-524 [FREE Full text] [doi: 10.1093/jamia/ocad220] [Medline: 37964688]
25. Rubinger L, Gazendam A, Ekhtiari S, Bhandari M. Machine learning and artificial intelligence in research and healthcare. Injury 2023 May;54 Suppl 3:S69-S73. [doi: 10.1016/j.injury.2022.01.046] [Medline: 35135685]

26. Pou-Prom C, Murray J, Kuzulugil S, Mamdani M, Verma AA. From compute to care: lessons learned from deploying an early warning system into clinical practice. Front Digit Health 2022 Sep 5;4:932123 [FREE Full text] [doi: 10.3389/fdgth.2022.932123] [Medline: 36133802]
27. Verma AA, Stukel TA, Colacci M, Bell S, Ailon J, Friedrich JO, et al. Clinical evaluation of a machine learning-based early warning system for patient deterioration. CMAJ 2024 Sep 15;196(30):E1027-E1037 [FREE Full text] [doi: 10.1503/cmaj.240132] [Medline: 39284602]
28. Sendak MP, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. JMIR Med Inform 2020 Jul 15;8(7):e15182 [FREE Full text] [doi: 10.2196/15182] [Medline: 32673244]
29. Brier GW. Verification of forecasts expressed in terms of probability. Mon Wea Rev 1950 Jan;78(1):1-3. [doi: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2]
30. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21(1):128-138. [doi: 10.1097/ede.0b013e3181c30fb2]
31. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. arXiv Preprint posted online on March 4, 2017 [FREE Full text] [doi: 10.5260/chara.21.2.8]
32. Moscovici M, Farrokhi F, Vangala L, Simpson AI, Kurdyak P, Jones RM. Violence risk prediction in mental health inpatient settings using the Dynamic Appraisal of Situational Aggression. Front Psychiatry 2024 Dec 10;15:1460332 [FREE Full text] [doi: 10.3389/fpsyt.2024.1460332] [Medline: 39720430]
33. Ogloff JR, Daffern M. The dynamic appraisal of situational aggression: an instrument to assess risk for imminent aggression in psychiatric inpatients. Behav Sci Law 2006 Dec 15;24(6):799-813. [doi: 10.1002/bsl.741] [Medline: 17171770]
34. Griffith JJ, Daffern M, Godber T. Examination of the predictive validity of the Dynamic Appraisal of Situational Aggression in two mental health units. Int J Ment Health Nurs 2013 Dec 30;22(6):485-492. [doi: 10.1111/inm.12011] [Medline: 23363378]
35. Ramesh T, Igoumenou A, Vazquez Montes M, Fazel S. Use of risk assessment instruments to predict violence in forensic psychiatric hospitals: a systematic review and meta-analysis. Eur Psychiatry 2018 Aug;52:47-53 [FREE Full text] [doi: 10.1016/j.eurpsy.2018.02.007] [Medline: 29626758]
36. Chu CM, Hoo E, Daffern M, Tan J. Assessing the risk of imminent aggression in institutionalized youth offenders using the dynamic appraisal of situational aggression. J Forens Psychiatry Psychol 2012 Apr 01;23(2):168-183 [FREE Full text] [doi: 10.1080/14789949.2012.668207] [Medline: 25999797]
37. Goldhill DR, White SA, Sumner A. Physiological values and procedures in the 24 h before ICU admission from the ward. Anaesthesia 1999 Jun 06;54(6):529-534 [FREE Full text] [doi: 10.1046/j.1365-2044.1999.00837.x] [Medline: 10403864]
38. Steitz BD, McCoy AB, Reese TJ, Liu S, Weavind L, Shipley K, et al. Development and validation of a machine learning algorithm using clinical pages to predict imminent clinical deterioration. J Gen Intern Med 2024 Jan 01;39(1):27-35 [FREE Full text] [doi: 10.1007/s11606-023-08349-3] [Medline: 37528252]
39. Hahn T, Nierenberg AA, Whitfield-Gabrieli S. Predictive analytics in mental health: applications, guidelines, challenges and perspectives. Mol Psychiatry 2017 Jan 15;22(1):37-43. [doi: 10.1038/mp.2016.201] [Medline: 27843153]
40. Hansen L, Bernstorff M, Enevoldsen K, Kolding S, Damgaard JG, Perfalk E, et al. Predicting diagnostic progression to schizophrenia or bipolar disorder via machine learning. JAMA Psychiatry 2025 May 01;82(5):459-469. [doi: 10.1001/jamapsychiatry.2024.4702] [Medline: 39969874]
41. Wolff J, Gary A, Jung D, Normann C, Kaier K, Binder H, et al. Predicting patient outcomes in psychiatric hospitals with routine data: a machine learning approach. BMC Med Inform Decis Mak 2020 Feb 06;20(1):21 [FREE Full text] [doi: 10.1186/s12911-020-1042-2] [Medline: 32028934]
42. Zhang BH, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018 Presented at: AIES '18; February 2-3, 2018; Orleans, LA URL: https://dl.acm.org/doi/10.1145/3278721.3278779 [doi: 10.1145/3278721.3278779]
43. Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev 2019 Jul 1;63(4/5):4:1-:15. [doi: 10.1147/JRD.2019.2942287]
44. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA URL: https://dl.acm.org/doi/10.5555/3295222.3295230

Abbreviations

AUC: area under the receiver operating characteristic curve
DASA-IV: Dynamic Appraisal of Situational Aggression–Inpatient Version
EMR: electronic medical record
EWS: early warning system
LSTM: long short-term memory

ML: machine learning
PRIME: Predictive Risk Identification for Mental Health Events

Edited by J Torous; submitted 29.Sep.2025; peer-reviewed by H Ryland, SCL Au; comments to author 17.Oct.2025; revised version received 10.Dec.2025; accepted 11.Dec.2025; published 06.Feb.2026.

Please cite as:
Dolatabadi E, Tamayo Velasquez V, Dabboussi AH, Wen D, Crawford J, Waddell AE, El Morr C
Advancing Psychiatric Safety With the Predictive Risk Identification for Mental Health Events Tool: Retrospective Cohort Study
JMIR Ment Health 2026;13:e84318
URL: https://mental.jmir.org/2026/1/e84318
doi:10.2196/84318
PMID:

Examining the Acceptability and Effectiveness of a Self-Directed, Web-Based Resource for Stress and Coping in University: Randomized Controlled Trial

Bilun Naz Böke[1], PhD; Jessica Mettler[2], PhD; Laurianne Bastien[1], MA; Sohyun Cho[1], MEd; Nancy Heath[1], PhD

[1]Department of Educational and Counselling Psychology, Faculty of Education, McGill University, 3700 McTavish Street, Montreal, QC, Canada

[2]Department of Psychology, Concordia University, Montreal, QC, Canada

Corresponding Author:
Bilun Naz Böke, PhD
Department of Educational and Counselling Psychology, Faculty of Education, McGill University, 3700 McTavish Street, Montreal, QC, Canada

Abstract

Background: University students face high levels of stress with limited support for coping and well-being. Campus mental health services are increasingly using digital resources to support students' stress management and coping capacity. However, the effectiveness of providing this support through web-based, self-directed means remains unclear.

Objective: Using a randomized controlled design, this study examined the acceptability and effectiveness of a self-directed, web-based resource containing evidence-based strategies for stress management and healthy coping for university students. The study additionally explored the potential benefits of screening and directing students to personalized resources aligned with their needs.

Methods: Participants consisted of 242 university students (193/242, 79.9% women; mean age 21.15 years) assigned to one of 3 groups (ie, automatically directed to personalized resources, nondirected, and waitlist comparison). They completed pre, post (4 wk), and follow-up (8 wk) measures for stress, coping, and well-being. The resource groups also completed acceptability measures at 2, 4, and 8 weeks after the web-based resource access.

Results: Results indicate high acceptability, reflecting students' satisfaction with the resource. Furthermore, significant decreases in stress and unhealthy coping, as well as significant increases in coping self-efficacy and healthy coping in the resource groups relative to the comparison group, were found. Interestingly, the directed approach showed no added benefit over nondirected resource access.

Conclusions: In summary, this study demonstrates the acceptability and effectiveness of a self-directed digital resource platform as a viable support option for university student stress and coping.

Trial Registration: ClinicalTrials.gov NCT07086001; https://clinicaltrials.gov/study/NCT07086001

(JMIR Ment Health 2026;13:e74205) doi:10.2196/74205

KEYWORDS

stress; coping; self-directed programming; web-based intervention; university students

Introduction

Background

University students consistently report high levels of stress and psychological distress and identify these as key factors that negatively impact their academic performance and engagement with their studies [1-4]. Supporting students in effectively coping with stress and distress is of critical importance to facilitate learning and development in university environments. To that end, technology-based approaches to delivering stress-management and well-being supports to university students have proliferated on campuses as supplemental means of supporting student stress management, coping capacity, and well-being [5]. Indeed, resources for students' self-directed use, such as websites, apps, or on-demand workshops, are increasingly popular given their benefits in improving access to support as well as the potential for reaching students who may be reluctant to seek other forms of mental health support or are on waiting lists for more specialized services [6]. In addition, the provision of resources for addressing stress and enhancing coping capacity is aligned with the recently proposed health theory of coping, which calls for enhancing the availability of evidence-based healthy coping strategies [7]. However, investigation into the acceptability, and even more critically, the effectiveness of digital, self-directed resources for nonclinical stress management and healthy coping support is limited. Thus, this study sought to explore the acceptability and effectiveness of a self-directed, web-based resource for enhancing students' stress management and coping capacity. Furthermore, the study also examined whether there would be any added benefit of screening students to assess stress and coping needs and then directing them to specific resources to match their needs for stress management and healthy coping support.

University Student Stress and Coping

University students' mental health and well-being have been a growing concern within higher education research and practice for many decades [8-10]. The most frequently identified factors impacting academic performance in recent population-level surveys (n=54,204) include stress (43.7%), anxiety (37.3%), depression (27.5%), and sleep difficulties (25.9%) [1]. Within Canada, university students (n=11,322) identified the same factors: stress (51.5%), anxiety (43.3%), depression (30.4%), and sleep difficulties (31.9%), as having had a negative impact on their academic performance over the past year [2]. For those pursuing a university education, this time in their lives often corresponds with their developmental transition to adulthood [11]. Coined in research literature as emerging adulthood, this developmental period is distinct from adulthood conceptually and as a subjective experience [12-15].

Emerging adulthood is a challenging yet unique time of exploration and settling into adult roles, often characterized as a time of feeling in-between [12,13]. While the transition to adulthood brings increased autonomy and responsibility, this period is also marked by instability across multiple life domains, including relationships, living arrangements, employment, and identity development. Navigating these changes can heighten vulnerability, stress, difficulties with coping, and mental health challenges. Notably, emerging adulthood is associated with elevated rates of engagement in risky and unhealthy coping behaviors in response to stress and distress [16-18]. For example, Böke et al [16] found that university students reporting higher stress were more likely to engage in substance use as a coping strategy. Conversely, research has shown that skill-based approaches to coping, such as problem-focused coping, defined as actively addressing the source of stress through problem-solving or planning, can buffer the negative impact of stress on well-being [19]. Taken together, there is a clear need to enhance access to evidence-based strategies and tools to support students in effectively managing stress and enhancing their capacity to cope with distress [19-22].

To enhance coping capacity among university students, understanding their decision-making processes in coping with stress is imperative. The health theory of coping offers a comprehensive framework for conceptualizing how students cope with stress and distress [7]. Stallman's health theory of coping considers all coping responses as adaptive, emphasizing their short-term efficacy in alleviating momentary stress or distress and further classifies coping responses into healthy and unhealthy coping behaviors based on the likelihood of adverse consequences. The theory presents a hierarchical model delineating coping responses across intensities, directly corresponding to the intensity of experienced stress or distress [7,19]. Low levels of stress or distress prompt low-intensity coping, encompassing both healthy (eg, positive self-talk, mindfulness, abdominal breathing) and unhealthy (eg, negative self-talk, cognitive rumination, suppression) responses. As distress intensifies, coping responses escalate, where higher intensity healthy strategies may include engaging in distracting activities, relaxation, physical exercise, or seeking social/professional support, while unhealthy responses may involve self-isolation, emotional eating, self-harm, substance use, or suicidality [7]. Acknowledging this hierarchical progression is pivotal in designing student support programs tailored to promote the availability of and engagement in evidence-based healthy coping behaviors.

Supporting Stress-Management and Building Coping Capacity

To date, efforts aimed at improving student mental health and well-being in university settings have included a wide variety of interventions targeting stress [23], depression [24], anxiety [6], resilience [25], and general mental health and well-being [26]. Increasingly, technology-based and digital tools (eg, websites, apps, chatbots, on-demand programming) are used with several systematic and meta-analytic reviews emphasizing the promise of the technology-based approach for improving key outcomes [5,26-28]. Furthermore, emerging research demonstrates the promise of sharing resources for students' self-directed use at their own pace and discretion [29-32].

For example, Fischer et al [33] demonstrated that self-directed interventions were effective in improving well-being and reducing stress, depression, and anxiety among both the general population and clinical samples when compared with active and inactive controls. This is supported by 2 meta-analytic reviews reporting significant effects of self-guided interventions for improving depressive symptoms in general population samples [34,35]. Among university students, a meta-analysis by Bolinski et al [29] found online mental health interventions (the majority were self-directed) to be effective for reducing anxiety and depression, although only a small and nonsignificant effect was reported for academic performance. In addition, Chung et al [36] examined the effectiveness of a university-wide, self-directed online mindfulness and well-being intervention and found improvements across stress, well-being, and mindfulness outcomes for those who engaged with the intervention over a duration of 3 or more weeks.

Self-directed or self-administered digital resources have the potential to serve as supplementary support for students and offer several advantages. First, they have the potential to reach those who may not access face-to-face services, who may not meet clinical criteria for specialized treatments, or are on waitlists for services, thus broadening access to evidence-based strategies and supports [30,35,37,38]. Second, the self-guided format is supportive of student autonomy and confidentiality as individuals can choose when, where, and how to access information and make use of resources most aligned with their individual needs [37]. Last, the web-based presentation of information and evidence-based strategies and techniques allows for a cost-effective, low-intensity, and adaptable (ie, possibility to update or change based on contextual needs) means to supplement existing mental health and well-being services on campus [6,39,40]. Furthermore, studies suggest that this modality is welcomed in universities [37,41] where up to 70% of students in a sample of 1224 indicated interest in self-guided mental health supports [42].

Issues With Supporting University Student Stress-Management and Healthy Coping

It should be noted that digital stress-management tools that are often developed for general adult or workplace populations and retroactively adapted for university students were found not to adequately address the developmental and contextual realities of this population [37]. As highlighted by Fleischmann et al [37], students face a unique combination of stressors at a precarious developmental transition, including academic and adjustment pressures, identity development, and unstable life circumstances that differ from those of working adults. Their findings underscore that students value support options that are specifically tailored to the academic context and their unique developmental needs while offering flexibility around fluctuating needs. Moreover, students report a desire for resources that reflect their lived experiences and offer personalized guidance and recommendations [37]. This suggests a need to include university students in the development of resources that are personalized to their unique needs, which may in turn enhance students' engagement with such resources. Despite emerging evidence of effectiveness for using digital, self-directed approaches to student support, research examining the effectiveness and acceptability of this approach is in its infancy. In addition, it is unclear to what extent digital, self-directed programming and resources are integrated into the university setting and used beyond their initial effectiveness trials [30]. Notably, even when interventions and programs for student mental health and well-being are shown to be effective, they are often only shared with students through the universities' health and wellness center, relying on students to proactively seek help to access these services. This poses a challenge because research consistently shows that university students exhibit low levels of help-seeking, leading to the underuse of many services and resources despite a high demand [43,44]. Additionally, earlier studies exploring means to support students' stress and coping have focused on addressing one aspect of stress or coping, such as mindfulness for stress, or breathing exercises for managing anxiety [45]. This signals a need for broader resources covering a wider array of topics and coping strategies to build coping capacity. Taken together, there is an urgent need to explore alternative approaches for resource delivery that facilitate students' universal and ongoing access to self-directed support options to comprehensively address stress and coping needs.

A persistent problem in university and a barrier to students' access to support is low rates of help-seeking, where stigma around mental health difficulties is considered to be a major contributor to students' reluctance to seek support [44,46]. Emerging research suggests that perceived mental health stigma can also contribute to students' responses to the format and modality of stress-management and well-being support delivery [47]. Specifically, Cho et al's [47] intervention study found that students' perceived mental health stigma did not impact their sustained satisfaction with a self-directed modality (ie, an infographic presenting evidence-based strategies for stress management and well-being), while it negatively impacted their sustained satisfaction with a live digital workshop presenting the same information with the presence of a facilitator. Beyond stigma, students may prefer digital, self-directed supports for several reasons, including concerns about confidentiality, social anxiety, and wanting to avoid social interactions focused on a topic that they would like to keep private. Overall, proactively connecting students to available resources is therefore an important consideration to navigate the effect of mental health stigma and other barriers on students' help-seeking behavior and promote their engagement with support services. One suggested solution for this is the use of brief screening measures to identify students' levels of need for support and recommend existing resources aligned with their personal needs [6,48,49]. Indeed, this approach has shown promise in clinical contexts as part of suicide prevention efforts in universities [49,50]. For example, in a large-scale study, Hasking et al [49] found that the use of a multivariable screener for suicidal risk followed by referral to a stepped telehealth intervention significantly increased resource use among university students classified as having the greatest need for intervention. Whether screening and tailoring resource recommendations can also promote students' engagement with, and use of, low-intensity stress-management and healthy coping resources in a nonclinical context remains to be explored.

Moreover, there is a need to consider students' uptake of stress-management and healthy coping strategies presented in self-directed resources. In a systematic review of prevention programs for stress, depression, and anxiety in university contexts, which included self-administered programming, Rith-Najarian et al [51] found inconsistencies in the assessment and reporting of information on uptake and adherence. Specifically, only 57% of the studies included in the review presented any information on adherence or completion, which prevented the authors from including adherence as a factor within their analyses [51]. A later study examining the effectiveness of a self-directed mindfulness intervention delivered over 12 weeks reported that students' access to the program modules peaked during the first 3 weeks, declined steeply over weeks 3 to 7, and then stabilized with a small increase in the final week 12 [36]. Overall, the authors reported that 58.7% of their total sample (n=833) did not access the mindfulness program at all over the duration of the semester-long study [36]. Assessing and reporting uptake or use of the provided resource is of particular importance in studies examining self-directed modalities where use can fluctuate over time and where the proportion of zero-uptake may be elevated. Furthermore, rates of uptake or use may influence the accuracy of effectiveness findings, and additional research is needed to better understand the relation between program uptake/adherence and outcomes of effectiveness [51].

This Study

In summary, despite the rapid proliferation of digital self-guided resources for university students, research examining the effectiveness of this approach for improving stress and coping is still in its infancy. Further research is needed to address gaps and deepen our understanding of what works best and how in the area of supporting university students' stress management and coping capacity [6,51]. Thus, using a randomized-controlled design, this study sought to examine the acceptability and effectiveness of a web-based, self-directed resource for university students containing evidence-based strategies for stress management and healthy coping. In addition, this study examined whether there would be any added benefit of using a screening approach to direct students to personalized resources aligned with their identified needs. Participants were randomly assigned to one of 3 groups: directed to personalized resources aligned with needs, nondirected but received all resources, or a waitlist comparison. Main outcomes assessed were participant ratings of acceptability, stress, coping (coping self-efficacy and coping behaviors), and well-being over time.

Specifically, the first objective (1) was to examine potential group differences (directed and nondirected resource groups only) in students' acceptability of the web-based resource over time. It was hypothesized that (H1) acceptability would be higher in the directed group when compared with the nondirected group over time. The second objective (2) was to examine the effectiveness of the digital self-directed resources in terms of group differences (directed, nondirected, and comparison) on outcome measures (ie, stress, coping, and well-being) and in terms of differences in scores over time between baseline, post, and follow-up measures. It is hypothesized (H2a) that the directed group will show greater improvements across stress, coping, and well-being outcomes over time than both the nondirected group and the comparison group. It is also hypothesized (H2b) that the nondirected group will show significant improvements across study outcomes relative to the comparison group. Last, the third objective (3) was to examine the effectiveness of the overall web-based, self-directed resource in terms of group differences (resource group; merged directed and nondirected vs the comparison group) on outcome measures and in terms of change in scores over time between baseline, post, and follow-up measures (ie, stress, coping, and well-being). It is hypothesized (H3) that the resource group will show significant improvements across study outcomes in relation to the comparison group.

Methods

Ethical Considerations

All procedures in this study were approved by the Research Ethics Board of McGill University (number 21-10-040). Informed consent was obtained prior to study participation; all participants were informed that they could choose to withdraw or end their participation in the study at any point without penalty or prejudice. Participant data have been aggregated for the purposes of data analysis and publication to respect privacy and confidentiality. Study participants received compensation of CAD $50 (US $36.40) via e-transfer for their participation. This study was registered as a randomized controlled trial on ClinicalTrials.gov (NCT07086001), and the associated study checklist is provided in Checklist 1.

Participants

Eligibility criteria included (1) being enrolled as a student at the university where the study took place and (2) being 18 years of age or older. Participants consisted of 242 university students recruited across a large university (193/242, 79.9% women; mean age 21.15). Participants were randomly assigned to one of 3 study groups (directed: 65/81 [80.5%] women, mean age 21.31; nondirected: 66/81, 81.5% women, mean age 21.07; comparison: 62/80, 77.8% women, mean age 21.06).

Resource Development and Content

The development of the web-based resource examined in this study was informed by 3 key foundational frameworks, namely, the health theory of coping [7], the theory of emerging adulthood [12,13], and Stepped Care 2.0 (SC2.0) [52,53]. Specifically, the health theory of coping provides a conceptual framework depicting university students' approaches to coping with stress and distress across a hierarchical spectrum where the intensity of the coping behavior is proportional to the intensity of experienced distress [7]. The theory of emerging adulthood and research describing general characteristics of this developmental period were instrumental in informing the topics and content developed and presented within the digital resource [12,13]. Last, SC2.0 presents a stepped, hierarchical framework for the organization of campus mental health care and services across incremental steps of intensity [52,53]. The resource tested within this study aligns with the lower intensity steps within SC2.0, and the framework has influenced and informed the screening and referral to personalized resources (ie, directed vs nondirected) model tested within this study. In addition, resource development followed a collaborative approach with a large team of university students (undergraduate and graduate), researchers, and university mental health service professionals consulting at each project stage (eg, conceptualization, material development, implementation, and data collection).

Overall, the theoretical foundations described above, the environmental scan of best practices in digital resource creation, as well as consultations with the project team informed the scope of topics and content areas to create research-informed resources with evidence-based strategies and tips. For example, students particularly requested resources for topics such as dealing with breakups, managing household responsibilities, managing stress around finances, setting and maintaining boundaries, and building social connections, among others. A priori, it was determined that resources would be presented in several multimedia formats (ie, text, audio, video, interactive infographic) to account for the diversity of preferences. In sum, there were over 50 different resources developed to highlight evidence-based strategies for healthy coping, addressing a broad scope of topics relevant to emerging adult university students in a demanding academic context. All resources were grouped in 5 main categories: managing stress, which presented strategies for coping with everyday stressors and enhancing emotion regulation capacity; enhancing performance, focused on skills around enhancing academic performance such as motivation, time management, and responding to academic setbacks; adulting addressed skills around the transitional life stage such as career exploration, relational changes (eg, breakups), and financial management; socializing offered guidance on building and maintaining meaningful social connections and dealing with loneliness; and well-being presented strategies that support psychological resilience such as gratitude, mindfulness, and self-awareness. Additionally, a psychoeducation and information-based section titled Understanding was created to share general statistics and information pertaining to university student stress, mental health, and well-being. The website also presented an additional resources section to connect students to, and encourage their use of, other services and resources they are eligible for at the university, in the local community, and through other websites and apps.

Importantly, given the web-based nature of the resource, accessibility of digital content was a key consideration throughout development and implementation. Consistent with Web Content Accessibility Guidelines 2.0 [54], features across content included accessible font styles and sizes, high-contrast color schemes, screen reader compatibility, plain language, and a mobile-optimized version of the website to support diverse user needs.

Procedure

Overview

Participants who expressed interest in participating in the study were asked to complete a brief digital demographics survey to facilitate their random assignment into the 3 different conditions within the study; namely, directed to resources based on reported need in the screening questionnaire (Group 1: directed), nondirected sharing of all resources (Group 2: nondirected), and waitlist comparison (Group 3: comparison). Participants were randomly assigned to the 3 conditions by the study lead author using IBM SPSS Statistics (version 23) tools, where participant IDs were randomly organized into 3 separate groups. Blinding was not deemed necessary as the study was conducted entirely digitally and directing to the resource was automated. Responses to the demographic questionnaire were used to ensure comparable samples across the different conditions in terms of participants' age, gender, and program of study. Following random assignment to the different conditions, all participants were asked to complete the baseline measures and the screening questionnaire (described in the measures section below). Although all participants were asked to complete the brief screening questionnaire, only those in the directed group subsequently received personalized instruction on how to use the resources and strategies provided in the digital resource.

Group 1 (Directed)

Immediately following the completion of the baseline survey, Group 1 was given access to the website presenting a collection of stress-management, motivation, healthy coping, well-being, and socializing resources. Additionally, based on their answers to the brief screener, Group 1 was directed to one of 3 unique pages on the website based on their responses on the screening questionnaire, demonstrating low, moderate, or high need for support around stress and coping. The directing process was automated using a scoring algorithm within the survey platform used in this study (ie, Qualtrics). Details on the screening questions, algorithm, and cut-off scores are provided in the Multimedia Appendix 1.

Group 2 (Nondirected)

Participants in Group 2 followed the same procedure as Group 1; however, they did not receive any personalized instruction and were simply directed to the home page of the website containing resources.

Group 3 (Comparison)

Participants in Group 3 constituted the waitlist comparison group. As such, they did not have access to any of the strategies hosted on the website during the data collection phase of the study. Participants in Group 3 were asked to complete web-based surveys identical to those completed by Groups 1 and 2. Although Group 3 did not have access to the strategies during the project, the full web-based resource was shared with the comparison group at the end of data collection.

In terms of data collection timeline, all groups completed measures (detailed in the next section) regarding their stress, coping, and well-being at the start of the study (baseline: T1), 4 weeks after the start of the study (post: T2), and 8 weeks following the start of the study (follow-up: T3). In addition, participants in Groups 1 and 2 completed a brief check-in to assess resource acceptability 2 weeks after baseline, which is when the resources were initially shared with participants.

Measures

Screening

The purpose of this screening questionnaire was to assess students' varying levels of need for support around stress, distress, coping, self-efficacy, loneliness, and social support to enable the directing of Group 1 (directed) to resources that match their need for stress-management and healthy coping support. This screener consisted of a 24-item researcher-designed measure comprised of a mix of single items assessing coping behaviors, financial stress, and access to community, as well as short versions of standardized measures that have been shown to be associated with university students' overall adjustment and well-being including, perceived stress [55], coping self-efficacy [56], loneliness [57], social support [58], and social connectedness [59]. Participants in the directed group were categorized as indicating high, moderate, or low need for stress-management and coping support based on their scores on the researcher-developed screening questionnaire and were subsequently directed to unique pages of the web-based resource. The scoring and categorization algorithm is described in the Multimedia Appendix 1. In brief, cut-off scores were set as the top/bottom 15th percentile score within the sample for each section of the screener; ie, stress and coping behaviors (general stress), perceived stress, coping self-efficacy (intrapersonal), loneliness, social support, and social connectedness (interpersonal). Participants with scores meeting or exceeding the cut-off across the general stress, interpersonal, and intrapersonal sections were categorized as having a high need for support. Participants with scores meeting or exceeding the cut-off in at least one section were categorized as having moderate need for support. Last, those with scores below the cut-off across all sections of the screener were categorized as having low need for support. The distribution of high, moderate, and low need categories is provided in Table 1. In terms of the pages they were directed to, those scoring in the high need category were directed to comprehensive resources for stress and coping support in the community, crisis lines, as well as specific help-seeking strategies (note that no participants in this study scored in the high need category across groups; thus, they did not receive the direction described above, and the implications are discussed in the Results and Discussion sections). Those indicating moderate need for support were directed to the full web-based resource and encouraged to use the presented strategies. Last, those indicating low need for support were directed to the understanding section of the website to provide further information around stress and coping, as well as a list of evidence-based stress-management and healthy coping strategies for their quick use in the event they feel a need.

Table . Participant demographic information and screener scores: full sample (n=212) and the subsample of participants (n=177) who reported at least some use of the digital resource.

	Full sample			Subsample		
	Directed	Nondirected	Comparison	Directed	Nondirected	Comparison
Age (years), mean (SD)	21.22 (2.68)	21.17 (3.13)	20.81 (2.19)	20.70 (1.79)	21.04 (3.21)	20.81 (2.19)
Gender, n (%)						
Woman	60 (83.3)	54 (81.8)	59 (79.7)	44 (81.5)	41 (83.7)	59 (79.7)
Man	11 (15.3)	9 (13.6)	14 (18.9)	10 (18.5)	7 (14.3)	14 (18.9)
Nonbinary	0 (0)	3 (4.5)	1 (1.4)	0 (0)	1 (2)	1 (1.4)
Prefer not to say	1 (1.4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Faculty of study, n (%)						
Agriculture & Environmental Science	5 (6.9)	6 (9.1)	5 (6.8)	5 (9.3)	3 (6.1)	5 (6.8)
Arts	18 (25)	17 (25.8)	21 (28.4)	12 (22.2)	13 (26.5)	21 (28.4)
Continuing Studies	1 (1.4)	0 (0)	0 (0)	1 (1.9)	0 (0)	0 (0)
Education	1 (1.4)	1 (1.5)	4 (5.4)	1 (1.9)	1 (2)	4 (5.4)
Engineering	1 (1.4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Law	5 (6.9)	4 (6.1)	2 (2.7)	3 (5.6)	2 (4.1)	2 (2.7)
Management	19 (26.4)	24 (36.4)	24 (32.4)	18 (33.3)	20 (40.8)	24 (32.4)
Medicine	1 (1.4)	2 (3)	3 (4.1)	1 (1.9)	2 (4.1)	3 (4.1)
Music	1 (1.4)	0 (0)	0 (0)	1 (1.9)	0 (0)	0 (0)
Nursing	3 (4.2)	1 (1.5)	0 (0)	2 (3.7)	1 (2)	0 (0)
Science	13 (18.1)	6 (9.1)	11 (14.9)	8 (14.8)	4 (8.2)	11 (14.9)
Other[a]	4 (5.6)	5 (7.6)	4 (5.4)	2 (3.7)	3 (6.1)	4 (5.4)
Screener score[b], n (%)						
Low need	44 (61.1)	42 (63.6)	50 (67.6)	34 (63)	29 (59.2)	50 (67.6)
Moderate need	28 (38.9)	24 (36.4)	24 (32.4)	20 (37)	20 (40.8)	24 (32.4)
High need	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

[a]The category of "Other" for Faculty of Study included those in cross-faculty programs (eg, Arts & Science).

[b]The scoring algorithm for the screener to determine low, moderate, and high need categories is provided in Multimedia Appendix 1.

Acceptability

Participants' ratings of the acceptability of the resources and strategies shared were assessed using a researcher-developed measure aligned with the Kirkpatrick New World Model for program evaluation [60]. Specifically, a total of 11 items assessed participants' (1) overall satisfaction with the resource (8 items; eg, "I found the website useful for me"; "The strategies presented in the website helped me better understand how to manage my stress and improve my wellness"; "I found that the website presented valuable strategies and techniques" rated on a 4-point Likert scale; 1="strongly disagree" to 4="strongly agree"), (2) frequency of actual and planned use of strategies (2 items; ie, "Over the past two weeks, how often did you use the strategies presented on the website?" and "Over the coming weeks, how often do you plan to use the strategies presented on the website?" Rated on a 4-point Likert scale; 1="every day" to 4="never") as well as (3) a single item to rate perceived impact for their well-being (ie, "Over the past two weeks, how would you rate the impact of the strategies presented on the website on your well-being?" Rated on a 4-point Likert scale; 1="no impact" to 4="high impact"). Scores were summed for the first part of the measure depicting satisfaction (ie, items 1 - 8), and the remaining items (actual and planned strategy use, impact on well-being) were analyzed as single-item responses. Internal consistency of the satisfaction subscale was good in this study (α=.88, .85, .87 at 2 weeks post baseline, T2, and T3, respectively). The complete version of the acceptability questionnaire is presented in the Multimedia Appendix 1.

Stress

Participants' perceived level of general stress was assessed using the 10-item version of the Perceived Stress Scale (PSS) [61]. This measure is a widely used self-report measure of adults' perception of stress. The items ask participants to indicate their experience of stress and the degree to which life situations are stressful on a 5-point scale; 0="never" to 4="very often." Items include statements such as "In the last two weeks, how often have you felt difficulties were piling up so high that you could not overcome them?" and "In the past two weeks, how often have you felt nervous and stressed?" Higher scores on the PSS represent greater perceived stress. The PSS has adequate internal reliability, construct validity, and predictive validity with reports of psychological and physical symptoms and the use of health services [55]. Although the original measure asks participants to report perceived stress over the last month, the measure was adapted in this study for consistency of timeline across measures; therefore, the prompt was adapted to ask that participants report their perceived stress over the past 2 weeks. Descriptive statistics for the PSS-10 in this study (Tables 2 and 3) were deemed comparable to those reported among other university student samples (mean 19.79, SD 6.37) [62]. The internal consistency of the PSS in this study was good (α=.83, .84, .85 at T1, T2, T3, respectively).

Table . Series of 3 (group: active, passive, comparison) × 3 (time: baseline, post, follow-up) mixed design ANOVAs for mental health and well-being outcomes among a subsample of participants who reported using the strategies presented in the digital resource (n=177).

Outcome	Time point	Directed (n=54), mean (SD)	Nondirected (n=49), mean (SD)	Comparison (n=74), mean (SD)
Stress				
Int[a,b] ($F_{3.807, 331.190}$=2.571, P=.04, η_p^2=.029)	Baseline	22.09 (5.68)	21.96 (6.24)	21.92 (6.20)
MET[c,d] ($F_{1.903, 331.190}$=6.613, P=.002, η_p^2=.037)	Post	20.77 (6.36)	19.70 (5.77)	21.40 (6.73)
MEG[e] ($F_{2, 174}$=0.770, P=.46, η_p^2=.009)	Follow-up	19.78 (5.46)	20.53 (6.35)	22.19 (6.92)
Coping self-efficacy				
Int ($F_{4, 348}$=2.395, P=.052, η_p^2=.027)	Baseline	143.10 (36.91)	136.66 (37.58)	143.78 (42.04)
MET[d] ($F_{2, 348}$=8.993, P<.001, η_p^2=.049)	Post	147.54 (42.91)	146.70 (37.24)	145.25 (45.41)
MEG ($F_{2, 174}$=.0325, P=.70, η_p^2=.004)	Follow-up	158.55 (38.30)	151.35 (39.45)	144.81 (43.78)
Healthy coping				
Int ($F_{4, 348}$=1.978, P=.098, η_p^2=.022)	Baseline	12.00 (3.94)	11.71 (3.11)	12.02 (3.56)
MET[d] ($F_{2, 348}$=15.962, P<.001, η_p^2=.084)	Post	13.12 (3.56)	12.66 (3.31)	12.34 (3.54)
MEG ($F_{2, 174}$=0.688, P=.50, η_p^2=.008)	Follow-up	13.69 (3.90)	13.46 (3.17)	12.46 (4.07)
Unhealthy coping				
Int[b] ($F_{3.697, 321.674}$=2.937, P=.02, η_p^2=.033)	Baseline	9.43 (9.43)	9.86 (3.15)	9.63 (3.93)
MET[d] ($F_{1.849, 321.674}$=9.603, P<.001, η_p^2=.052)	Post	8.67 (8.67)	9.52 (3.01)	9.54 (3.57)
MEG ($F_{2, 174}$=1.235, P=.29, η_p^2=.014)	Follow-up	8.20 (8.20)	8.19 (2.98)	9.57 (3.60)
Well-being				
Int ($F_{4, 348}$=0.611, P=.65, η_p^2=.007)	Baseline	3.09 (0.65)	3.16 (0.59)	3.24 (0.64)
MET ($F_{2, 348}$=0.762, P=.46, η_p^2=.004)	Post	3.17 (0.66)	3.23 (0.59)	3.24 (0.77)
MEG ($F_{1, 174}$=0.169, P=.84, η_p^2=.002)	Follow-up	3.18 (0.66)	3.19 (0.67)	3.19 (0.72)

[a]Int: Interaction.

[b] $P<.05$.

[c] MET: main effect of time.

[d] $P<.001$; Bonferroni correction ($P=.05/3=.0167$) was used at the level of main effects to account for multiple comparisons.

[e] MEG: main effect of group.

Table . Series of 2 (group: resource, comparison) × 3 (time: baseline, post, follow-up) mixed design ANOVAs for mental health and well-being outcomes after merging the directed and nondirected groups into a single resource group (n=177).

Outcome	Time point	Resource group, mean (SD)	Comparison, mean (SD)
Stress			
Int[a,b] ($F_{1.911, 334.382}$=3.597, P=.03, η_p^2=.020)	Baseline	22.03 (5.92)	21.92 (6.20)
MET[b,c] ($F_{1.911, 334.382}$=4.230, P=.02, η_p^2=.024)	Post	20.26 (6.08)	21.40 (6.73)
MEG[d] ($F_{1, 175}$=1.530, P=.22, η_p^2=.009)	Follow-up	20.14 (5.89)	22.19 (6.92)
Coping self-efficacy			
Int[b] ($F_{2, 350}$=4.196, P=.02, η_p^2=.023)	Baseline	140.04 (37.19)	143.78 (42.04)
MET[e] ($F_{1.943, 339.997}$=5.448, P=.005, η_p^2=.030)	Post	147.14 (40.12)	145.25 (45.41)
MEG ($F_{1, 175}$=0.257, P=.61, η_p^2=.001)	Follow-up	155.12 (38.83)	144.81 (43.78)
Healthy coping			
Int[b] ($F_{2, 350}$=3.894, P=.02, η_p^2=.022)	Baseline	11.86 (3.55)	12.02 (3.56)
MET[e] ($F_{2, 350}$=11.259, P<.001, η_p^2=.060)	Post	12.90 (3.43)	12.34 (3.54)
MEG ($F_{1, 175}$=1.109, P=.29, η_p^2=.006)	Follow-up	13.58 (3.56)	12.46 (4.07)
Unhealthy coping			
Int[b] ($F_{1.854, 324.520}$=4.784, P=.01, η_p^2=.027)	Baseline	9.63 (3.11)	9.63 (3.93)
MET[e] ($F_{1.854, 324.520}$=5.532, P=.005, η_p^2=.031)	Post	9.08 (2.96)	9.54 (3.57)
MEG[d] ($F_{1, 175}$=1.921, P=.17, η_p^2=.011)	Follow-up	8.20 (3.13)	9.57 (3.60)
Well-being			
Int[a] ($F_{2, 350}$=.989, P=.37, η_p^2=.006)	Baseline	3.13 (0.62)	3.24 (0.64)
MET[c] ($F_{2, 350}$=.513, P=.60, η_p^2=.003)	Post	3.20 (0.63)	3.24 (0.77)
MEG[d] ($F_{1, 175}$=.367, P=.55, η_p^2=.002)	Follow-up	3.18 (0.66)	3.19 (0.72)

[a]Int: Interaction.
[b]*P*<.05.
[c]MET: main effect of time.
[d]MEG: main effect of group.
[e]*P*<.001; Bonferroni correction (*P*=.05/3=.0167) was used at the level of main effects to account for multiple comparisons.

Coping

Participants' belief in their ability to cope with general difficulty and distress was assessed using the Coping Self-Efficacy Scale (CSE) [56]. The CSE is a measure of one's confidence in effectively engaging in coping behaviors in the face of challenges. There are 26 items and 3 subscales within the CSE; namely, problem-focused coping (12 items), emotion-focused coping (9 items), and social support (5 items). Participants are asked to rate their confidence in their ability to perform the listed coping behaviors (eg, "find solutions to your most difficult problems," "see things from the other person's point of view during a heated argument") on an 11-point Likert scale; 0="cannot do at all" to 10="certainly can do." Higher scores on the CSE represent greater belief in one's own ability to cope with difficulty. The CSE demonstrated negative correlations with perceived stress, burnout [56], and emotion regulation difficulties [63]. Conversely, the CSE is positively correlated with optimism [56]. In this study, the prompt for this measure was adapted to ask participants about their confidence in their ability to perform the listed coping behaviors, specifically over the past 2 weeks, and the internal consistency of the full CSE was excellent (α=.93, .95, and .95, at T1, T2, and T3, respectively).

In addition, the Coping Index (CI) [64] was used to assess students' engagement in healthy and unhealthy coping behaviors over the duration of the study. The CI is a 20-item measure of engagement with healthy (10 items) and unhealthy (10 items) coping behaviors, which are aligned with the health theory of coping framework [7]. The measure consists of items that list common healthy and unhealthy coping behaviors, such as "talk things over with family or friends," "do relaxing activities," or "have negative self-talk." Participants are asked to indicate how often they engage in each behavior listed when they feel stressed or distressed on a 4-point Likert scale (0= "I don't do this at all" to 3= "I do this most of the time"). Higher scores on the healthy coping subscale indicate greater frequency of engagement in healthy coping behaviors; similarly, higher scores on the unhealthy coping subscale indicate greater frequency of engagement in unhealthy coping in response to stress or distress. This measure has been found to have satisfactory test-retest reliability in previous studies (α=.71) [65]. In this study, internal consistency of the healthy coping subscale was poor (α=.57, .57, .64 at T1, T2, T3, respectively), and the unhealthy coping subscale was also poor (α=.53, .53, .58 at T1, T2, T3, respectively). This is expected and deemed borderline acceptable for research purposes [66], given that the items within the subscales of the CI assess unique coping behaviors that may not necessarily have high agreement between them.

Well-Being

Well-being was assessed using the Warwick-Edinburgh Mental Well-Being Scale (WEMWBS) [67]. This measure consists of 14 positively worded items assessing overall subjective well-being. Participants are asked to rate statements such as "I've been feeling good about myself" according to their experience over the past 2 weeks on a 5-point Likert scale (1="none of the time" to 5="all of the time"). A higher WEMWBS score represents a higher level of mental well-being. The WEMWBS has demonstrated good internal consistency within university students (α=.89) and general population samples (α=.91). Test-retest reliability after a one-week delay was also high (α=.83) [67]. The internal consistency of the WEMWBS in this study was excellent (α=.91, .92, .93 at T1, T2, T3, respectively).

Data Analytic Plan

The overarching purpose of the study was to examine the acceptability and the effectiveness of a self-guided digital resource for university student stress, coping, and well-being outcomes. Preliminary analyses (ie, a one-way ANOVA, chi-square tests) were conducted to ensure comparability of the 3 study groups on demographic variables such as age, gender, and faculty of study at baseline. Given the importance of actual engagement with the digital resource for the accurate assessment of acceptability [51], the analyses of acceptability (Objective 1) were conducted both within the full study sample and a subsample of participants consisting of those who reported using the resources at least sometimes across all timepoints. Preliminary descriptive statistics were computed to examine students' satisfaction with the digital resource, their reported and intended use of strategies, and the perceived impact of using the strategies on their well-being among both the directed and nondirected groups. Group differences in satisfaction and strategy use ratings were examined using a series of 2-way mixed design ANOVAs to examine the effects of condition (directed vs nondirected delivery of resources) and time (baseline, post, and follow-up) on student ratings of satisfaction and strategy use, as well as the reported impact of strategy use on their well-being. Across all analyses, the Bonferroni correction was used across at the level of main effects, simple main effects, and pairwise comparisons to account for multiple comparisons.

Notably, there were a total of 35 (14.46% of the total sample) participants (mean age 22.00, SD 3.48; 78.9% women) in the resource groups that reported never using the presented digital resource and strategies. In a resource evaluation study, those who were assigned to a resource group but chose not to engage with the resource cannot comment on the resources, nor would we expect the resources to effect a change, and this data may interfere with the accurate evaluation of the effectiveness of the resources. Compared with students who reported using the strategies (n=103; n=177 when including 74 participants from the comparison group), those who reported never using the strategies (n=35) were not significantly different on any of the study variables (stress, coping, and well-being) at baseline. Therefore, those who reported never using the strategies were excluded from the subsequent analyses, which were only conducted among the subsample of participants who reported using the resource at least sometimes across the 3 timepoints (directed: n=54, mean age 20.70, SD 1.79, 81.5% women; nondirected: n=49, mean age 21.04, SD 3.208, 83.7% women; comparison: n=74, mean age 20.81, SD 2.19, 79.7% women). The criterion of "at least sometimes" was used to ensure that participants had some degree of exposure to the web-based resource, as prior research suggests that even minimal engagement is necessary for participants to provide informed feedback on effectiveness and acceptability [51]. This threshold was therefore chosen to distinguish between no use and at least some use of the web-based resource being tested.

Thus, for the accurate assessment of effectiveness (Objective 2), analyses were restricted to the subsample consisting of participants who reported at least some use of the digital resource across the study timeline. A series of 3 (condition: directed, nondirected, waitlist comparison) × 3 (time: baseline, post, follow-up) mixed-design ANOVAs were used to examine potential changes in stress, coping, and well-being over time.

Last for objective 3, which sought to examine the overall effectiveness of the digital resource against a business-as-usual comparison group, the directed and nondirected groups were merged into one "resource group" to facilitate this analysis. A series of 2 (condition: resource group, waitlist comparison) × 3 (time: baseline, post, follow-up) mixed-design ANOVAs were used to examine potential changes in stress, coping, and well-being over time. Across all analyses, follow-up examination of main effects and simple main effects of group and time was conducted to locate any observed differences by group or over time. Bonferroni corrections were used across main effects and simple main effects analyses to account for multiple comparisons. IBM SPSS (version 23; IBM Corp) was used for all analyses in this study.

Results

Preliminary Analyses

Participants were randomly assigned to the directed, nondirected, and comparison groups following their completion of the demographic questionnaire. A one-way ANOVA revealed no differences based on age across the study groups, $F_{2,229}$=0.139, P=.87. Two chi-square tests of independence revealed no associations across the groups by gender, χ^2_6=5.9, P=.44, or faculty of study, χ^2_{22}=18.3, P=.69. Thus, the efficacy of the randomization and comparability of the study groups was supported. A total of 19 participants were excluded from all analyses, given that most of their digital surveys were incomplete. Missing values analyses demonstrated less than 5% of missing data within each timepoint and group, which were imputed using the Expectation Maximization method. There were 4 univariate outliers identified (z>|3.29|) which were winsorized for data conservation. Thus, the final study sample consisted of 212 participants (mean age 21.06, SD 2.67, 81.6% women). As noted above, this study also considered the subsample of participants who reported at least some use of the strategies shared on the web-based resource (Figure 1 displays the participant flow diagram). Demographic characteristics and screener scores of both the full sample and the subsample of participants are displayed in Table 1. Interestingly, participants' scores on the screener indicate either low or moderate need for stress-management and healthy coping support, with no participant scores signaling high need. The proportion of low versus moderate need, as indicated by screener scores, was comparable across all study groups (directed, nondirected, and comparison).

Figure 1. Participant flow diagram.

Objective 1: Acceptability of the Self-Directed Digital Resource as Assessed by Group Differences (Directed Vs Nondirected) Over Time (Baseline to Follow-up) on Overall Resource Satisfaction, Actual and Planned Strategy Use, and Perceived Impact on Well-Being

Participants in both the directed (Group 1) and nondirected (Group 2) conditions rated the digital resource very highly, with specific ratings across the acceptability questionnaire for each group across time depicted in the Multimedia Appendix 1. Overall, participants indicated that the strategies presented in the digital resource were valuable (90% and 92% agreed in Groups 1 and 2, respectively), presented in an engaging manner (83% and 86% in Groups 1 and 2, respectively), and easy to understand (93% and 94% in Groups 1 and 2, respectively). Similarly, up to 83% of those in the directed group and 79% of those in the nondirected group agreed that the strategies presented helped them better understand how to manage their stress and improve their wellness.

A 2-way mixed design ANOVA to assess group differences over time for overall satisfaction with the digital resource (sum score of acceptability items 1 to 8) revealed no significant group by time interaction; $F_{1.764,206.390}$=0.015, P=.98, η_p^2=.000 (Table 4). Similarly, no interactions were found for strategy use, $F_{1.793,208.039}$=.204, P=.79, η_p^2=.002; planned strategy use, $F_{2,232}$=1.554, P=.21, η_p^2=.013; and perceived impact of strategy use on well-being; $F_{2,234}$=0.067, P=.93, η_p^2=.001. Analyses of main effects revealed no significant changes in strategy use over time using the Bonferroni correction; $F_{1.793,208.039}$=3.576, P=.03, η_p^2=.030. Similarly, there was no significant main effect of time (MET) for participants' ratings of perceived impact of strategy use on their well-being; $F_{2,234}$=3.694, P=.03, η_p^2p2=.031. Thus, the first hypothesis (H1), expecting higher overall acceptability (satisfaction, strategy use, and impact on well-being) within the directed group, was not supported, with both groups reporting comparably high levels of acceptability for the digital resource.

Table . Series of 2 (group: directed, nondirected) × 3 (time: pre, post, follow-up) mixed design ANOVAs for acceptability of web-based resource.

Sample and outcome	Time point	Directed, mean (SD)	Nondirected, mean (SD)
Full sample (directed: n=54; nondirected: n=49)			
Satisfaction sum			
Int[a] ($F_{1.764, 206.390}$=0.015, *P*=.98, η_p^2=.000)	Pre	23.60 (4.57)	23.91 (3.02)
MET[b] ($F_{1.764, 206.390}$=1.332, *P*=.27, η_p^2=.011)	Post	23.52 (3.85)	23.77 (3.13)
MEG[c] ($F_{1, 117}$=0.176, *P*=.68, η_p^2=.002)	Follow-up	24.03 (4.00)	24.23 (3.87)
Strategy use			
Int[a] ($F_{1.793, 208.039}$=0.204, *P*=.79, η_p^2=.002)	Pre	3.10 (0.50)	3.04 (0.51)
MET[d] ($F_{1.793, 208.039}$=3.576, *P*=.03, η_p^2=.030	Post	3.05 (0.52)	3.05 (0.52)
MEG[c] ($F_{1, 116}$=0.097, *P*=.76, η_p^2=.001)	Follow-up	2.95 (0.68)	2.93 (0.54)
Planned strategy use			
Int[a] ($F_{2, 232}$=1.554, *P*=0.21, η_p^2=.013)	Pre	2.76 (0.56)	2.61 (0.65)
MET[b] ($F_{2, 232}$=1.696, *P*=.19, η_p^2=.014)	Post	2.61 (0.66)	2.66 (0.58)
MEG[c] ($F_{1, 116}$=0.479, *P*=.49, η_p^2=.004)	Follow-up	2.79 (0.68)	2.70 (0.63)
Impact on well-being			
Int[a] ($F_{2, 234}$=0.067, *P*=.93, η_p^2=.001)	Pre	2.44 (0.82)	2.39 (0.76)
MET[b] ($F_{2, 234}$=3.694, *P*=.03, η_p^2=.031)	Post	2.54 (0.78)	2.48 (0.81)
MEG[c] ($F_{1, 117}$=0.317, *P*=.58, η_p^2=.003)	Follow-up	2.65 (0.83)	2.55 (0.74)
Subsample (directed: n=46; nondirected: n=41)			
Satisfaction sum			
Int[a] ($F_{1.696, 144.162}$=0.266, *P*=.73, η_p^2=.003)	Pre	24.57 (4.08)	24.83 (2.62)
MET[b] ($F_{1.696, 144.162}$=1.894, *P*=.16, η_p^2=.022)	Post	24.41 (3.12)	24.63 (2.89)
MEG[c]($F_{1, 85}$=0.032, *P*=.86, η_p^2=.000)	Follow-up	25.26 (3.14)	25.07 (2.92)
Strategy use			

Sample and outcome	Time point	Directed, mean (SD)	Nondirected, mean (SD)
Full sample (directed: n=54; nondirected: n=49)			
Int[a] ($F_{1.610, 135.257}$=0.479, P=.58, η_p^2=.006)	Pre	2.89 (0.31)	2.85 (0.36)
MET[b] ($F_{1.610, 135.257}$=3.447, P=.04, η_p^2=.039)	Post	2.85 (0.36)	2.85 (0.36)
MEG[c] ($F_{1, 84}$=.009, P=.92, η_p^2=.000)	Follow-up	2.72 (0.54)	2.78 (0.42)
Planned strategy use			
Int[a] ($F_{2, 168}$=1.810, P=.17, η_p^2=.021)	Pre	2.62 (0.49)	2.44 (0.63)
MET[b] ($F_{2, 168}$=1.343, P=.26, η_p^2=.016	Post	2.49 (0.66)	2.54 (0.55)
MEG[c] ($F_{1, 84}$=0.108, P=.74, η_p^2=.001)	Follow-up	2.60 (0.62)	2.63 (0.62)
Impact on well-being			
Int[a] ($F_{2, 170}$=0.665, P=.51, η_p^2=.008)	Pre	2.70 (0.66)	2.66 (0.57)
MET[b] ($F_{2, 170}$=5.299, P=.01, η_p^2=.059)	Post	2.83 (0.57)	2.78 (0.57)
MEG[c] ($F_{1, 85}$=0.811, P=.37, η_p^2=.009)	Follow-up	2.98 (0.49)	2.80 (0.51)

[a]Int: Interaction.

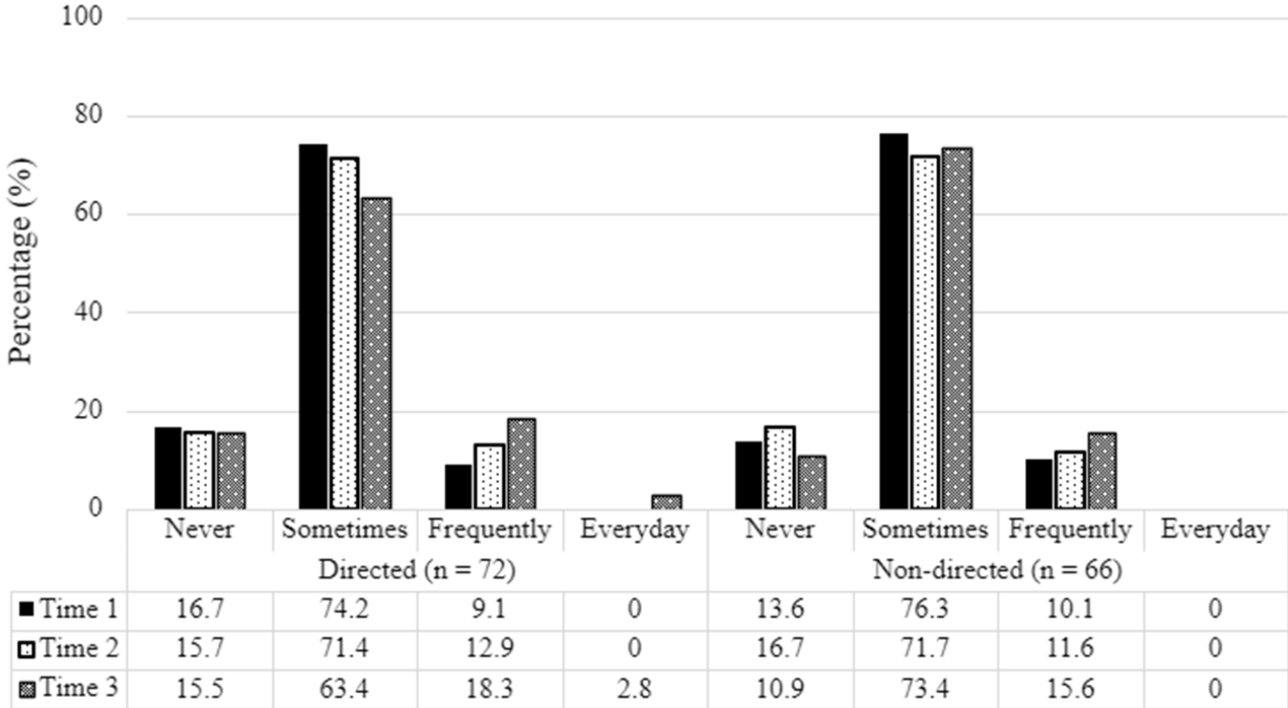
[b]MET: main effect of time.

[c]MEG: main effect of group.

[d]P<.05, Bonferroni correction (P=.05/2=.025) was used at the level of main effects to account for multiple comparisons.

Given the importance of strategy and resource use for the accurate assessment of acceptability and effectiveness (Figure 2), the same analyses were repeated among the subsample of participants who reported using the strategies presented in the web-based resource at least sometimes across all 3 timepoints (baseline to follow-up). Results revealed no statistically significant group by time interaction for overall satisfaction; $F_{1.696,144.162}$=0.266, P=.73, η_p^2=.003, strategy use; $F_{1.610,135.257}$=0.479, P=.58, η_p^2=.006, planned strategy use; $F_{2,168}$=1.810, P=.17, η_p^2=.021, and perceived impact on well-being; $F_{2,170}$=0.665, P=.51, η_p^2=.008 (Table 4). Examination of main effects revealed no significant changes in strategy use over time for both groups using the Bonferroni correction; $F_{1.793,135.257}$=0.479, P=.04, η_p^2=.039. Impact on well-being also did not change over time for both the directed and nondirected groups; $F_{2,170}$=5.299, P=.01, η_p^2=.059. Overall, contrary to the first hypothesis (H1), the directed and nondirected groups did not differ in terms of overall resource acceptability, strategy use, plan for strategy use, and reported impact of strategy use on well-being.

Figure 2. Percentage of participants reporting strategy use frequency in directed and nondirected groups, illustrating adherence to strategy use across Time 1, Time 2, and Time 3.



	Directed (n = 72)				Non-directed (n = 66)			
	Never	Sometimes	Frequently	Everyday	Never	Sometimes	Frequently	Everyday
Time 1	16.7	74.2	9.1	0	13.6	76.3	10.1	0
Time 2	15.7	71.4	12.9	0	16.7	71.7	11.6	0
Time 3	15.5	63.4	18.3	2.8	10.9	73.4	15.6	0

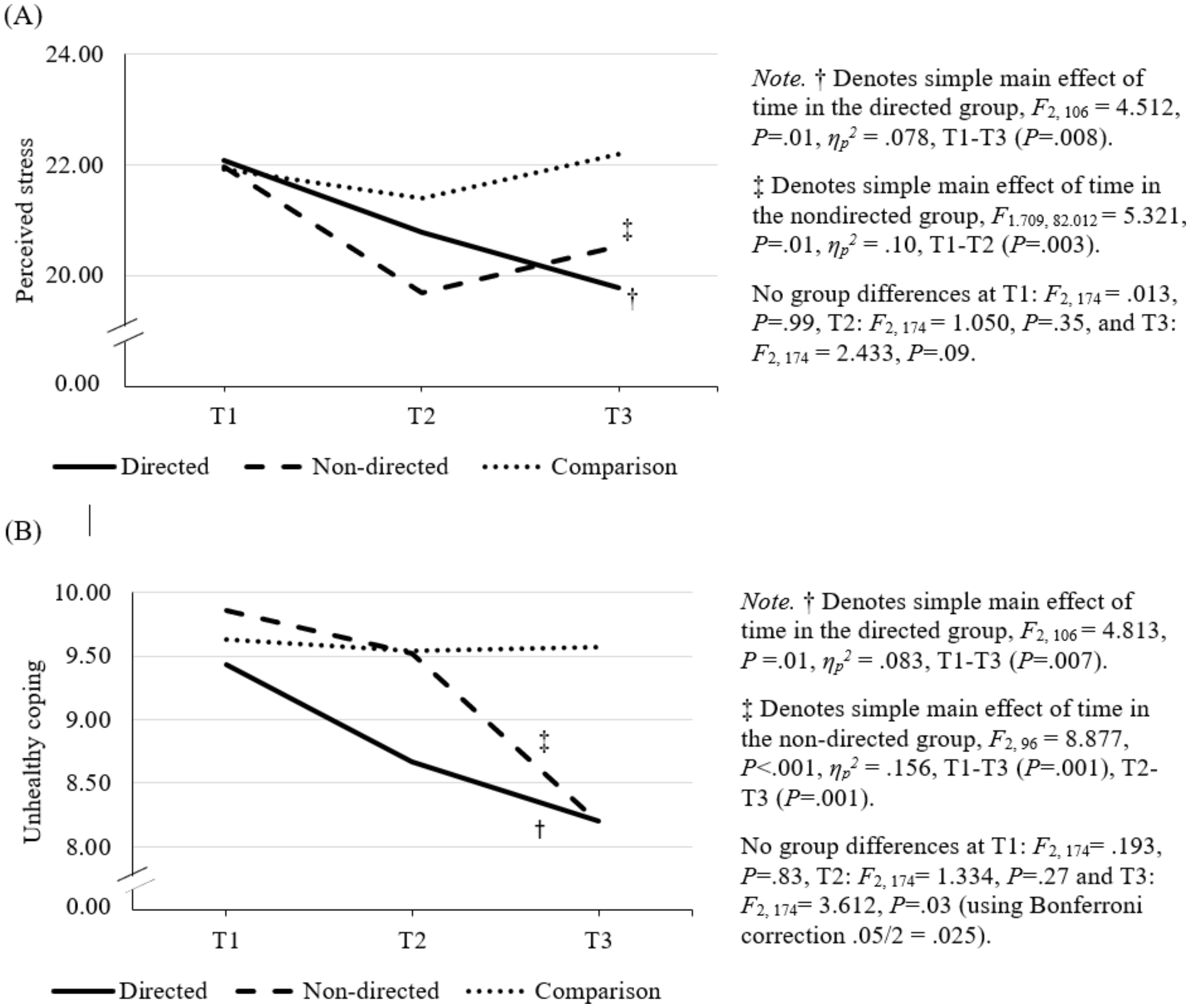
Objective 2: Effectiveness of the Self-Directed Digital Resource as Assessed by Group Differences (Directed Vs Nondirected Vs Comparison) Over Time (Baseline, Post, and Follow-Up) on Stress, Coping, and Well-Being Outcomes

A series of 2-way mixed design ANOVAs was conducted to assess group (directed, nondirected, and comparison) by time (baseline; T1, post; T2, follow-up; T3) interactions for stress, coping (coping self-efficacy, healthy coping, unhealthy coping behaviors), and well-being outcomes. As depicted in Table 2 and Figure 3, results revealed significant group-by-time interactions for stress and unhealthy coping; however, no significant interactions were found for coping self-efficacy, healthy coping, or well-being. Partially supporting hypothesis H2a, the directed group demonstrated significant improvements across stress and unhealthy coping in contrast to the comparison group; however, there were no differences between the directed and nondirected groups. Hypothesis H2b pertaining to changes in stress, coping, and well-being in the directed group relative to the comparison group was also partially supported.

Examination of simple main effects of group using the Bonferroni correction revealed no differences between groups for either stress or unhealthy coping across any of the timepoints. Patterns for the simple MET indicate that stress (*P*=.01, $_p^2$=.078) and unhealthy coping (*P*=.01, $_p^2$=.10) decreased over time within both the directed and nondirected groups but stayed stable across timepoints within the comparison group (Figure 3). Specifically, the observed decrease in stress took place between T1 and T3 (*P*=.008) for the directed group, and between T1 and T2 (*P*=.003) for the nondirected group. Unhealthy coping decreased between T1 and T3 in both groups (directed: *P*=.007, nondirected: *P*=.001), and the decrease between T2 and T3 (*P*=.001) was significant for the nondirected group.

Analyses of main effects for the nonsignificant interactions revealed a significant MET for coping self-efficacy (*P*<.001, $_p^2$=.049) and healthy coping (*P*<.001, $_p^2$p2=.084) with pairwise comparisons using the Bonferroni correction revealing a significant increase in coping self-efficacy from T1 to T3 (*P*<.001). Similarly, healthy coping showed a significant increase from T1 to T2 (*P*=.001) and from T1 to T3 (*P*<.001) across all groups.

Figure 3. Scores on perceived stress and unhealthy coping by time and group (directed, nondirected, and comparison), depicting simple main effects and pairwise comparisons for each outcome.

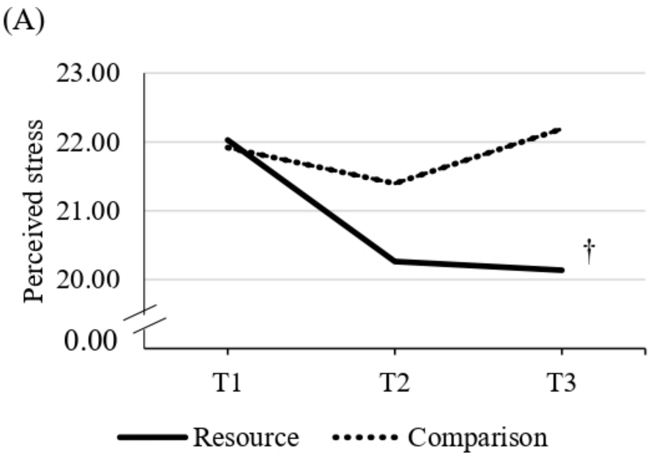


Objective 3 (Merged Groups): Effectiveness of the Self-Directed Digital Resource as Assessed by Group Differences (Resource Group Vs Comparison) Over Time (Baseline, Post, and Follow-Up) on Stress, Coping, and Well-Being Outcomes

A series of 2-way mixed design ANOVAs was conducted to assess group (resource group; merged directed and nondirected vs comparison) by time (baseline; T1, post; T2, follow-up; T3) interactions for stress, coping (coping self-efficacy, healthy coping, unhealthy coping behaviors), and well-being outcomes. As depicted in Table 3, significant group-by-time interactions were found for stress and coping outcomes, although no interaction was detected for well-being. As expected, results revealed significant decreases in stress and unhealthy coping, as well as increases in coping self-efficacy and healthy coping among the resource group over time, in contrast to the comparison group. Thus, hypothesis H3 was partially supported, given that no changes in well-being were detected.

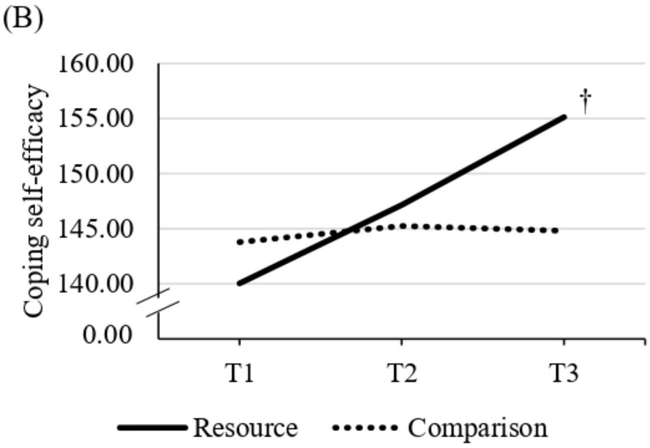
Analyses of simple main effects of time and group for the outcomes of stress, coping self-efficacy, healthy, and unhealthy coping are presented in Figure 4. In terms of the simple main effects of time, the resource group showed significant decreases in stress (P=.001, $_p^2$=.073) and unhealthy coping (P<.001, $_p^2$=.110), and significant increases in coping self-efficacy (P<.001, $_p^2$=.087) and healthy coping (P<.001, $_p^2$=.133) over time, in contrast to the comparison group. The observed changes over time took place between T1 and T3 for all outcomes (Figure 4), with significant changes detected between T1 and T2 for stress (decrease; P=.001) and healthy coping (increase; P=.002). Furthermore, coping self-efficacy significantly increased (P=.01) and unhealthy coping decreased (P=.002) between T2 and T3 within the resource group. In terms of the simple main effect of group, the resource group reported significantly lower unhealthy coping (P=.008, $_p^2$=.04) at the follow-up timepoint in contrast to the comparison group; no other group differences were detected between the resource and comparison groups.

Figure 4. Scores on stress and coping outcomes by time and group (resource, comparison), depicting simple main effects of time and group as well as pairwise comparisons for each outcome.



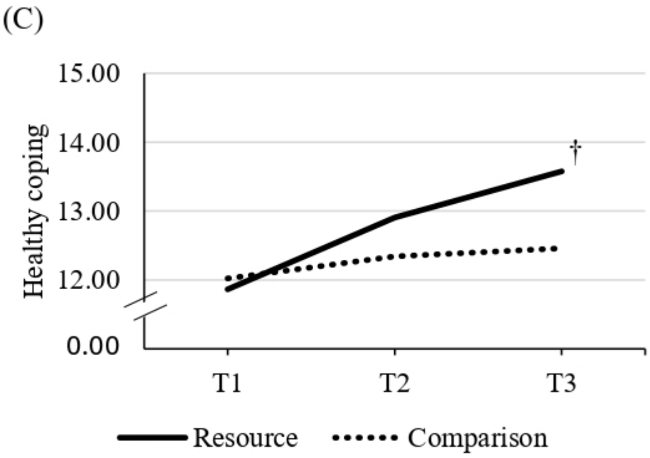
Note. † Denotes simple main effect of time in the resource group, $F_{1.882, 15128}$=8.084, P=.001, η_p^2 = .073, T1-T2 (P=.001), T1-T3 (P=.002).

No simple main effect of group was detected at T1: $F_{1, 175}$= .014, P=.906, T2: $F_{1, 175}$= 1.371, P=.243, and T3: $F_{1, 175}$= 4.517, P=.035, η_p^2 = .025.



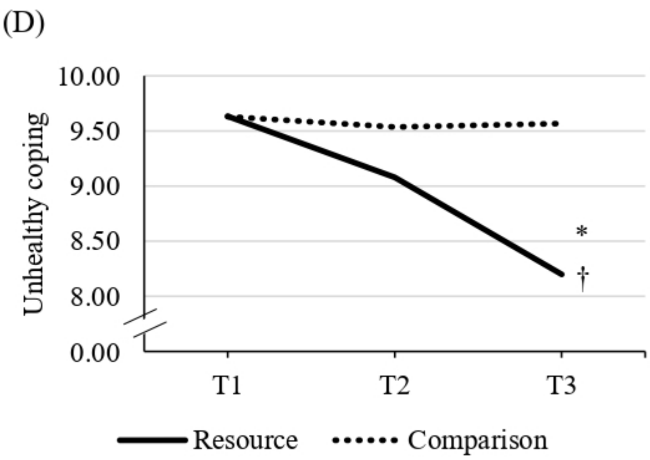
Note. † Denotes simple main effect of time in the resource group, $F_{2, 204}$= 9.683, P<.001, η_p^2 = .087, T1-T3 (P<.001), T2-T3 (P=.014).

No simple main effect of group was detected at T1: $F_{1, 175}$=0.391, P=.53, T2: $F_{1, 175}$= .086, P=.77, and T3: $F_{1, 175}$= 2.728, P=.10.



Note. † Denotes simple main effect of time in the resource group, $F_{2, 204}$= 15.585, P<.001, η_p^2 = .133, T1-T2 (P=.002), T1-T3 (P<.001).

No simple main effect of group was detected at T1: $F_{1, 175}$=0.085, P=.77, T2: $F_{1, 175}$= 1.133, P=.29, and T3: $F_{1, 175}$= 3.755, P=.05.



Note. † Denotes simple main effect of time in the resource group, $F_{1.886, 192.406}$= 12.552, P<.001, η_p^2 = .110, T1-T3 (P<.001), T2-T3 (P=.002).

* Denotes simple main effect of group at T3: $F_{1, 175}$= 7.266, P=.008, η_p^2 = .04. No simple main effect of group was detected at T1: $F_{1, 175}$= .000, p=.995, or T2: $F_{1, 175}$= .866, P=.35, and T2: $F_{2, 175}$= 3.755, P=.05.

Discussion

Principal Findings

This study sought to examine the acceptability and effectiveness of sharing a collection of evidence-based stress-management and healthy coping strategies and multimedia resources on a website for university students' self-directed use. Overall, students rated the resources and strategies presented on the website very highly, with comparably high rates of satisfaction reported by both those who received personalized recommendations after screening (ie, directed) and those who did not receive personalized recommendations (ie, nondirected). This finding is consistent with previous studies reporting high levels of receptivity and interest for digital, self-directed support options among university populations [32,42,68]. However, it was interesting that there was no added benefit of the screening and sharing personalized recommendations approach within this study. It is possible that high satisfaction with the overall web-based resource and the breadth of information shared constitutes a ceiling effect that prevented the detection of any unique benefits of screening in this study. This is consistent with previous findings where university students reported high levels of satisfaction with a self-directed, video outreach program [69]. These results potentially allude to students' high receptivity to information about stress management and healthy coping that is presented in multimedia, self-paced, and visually engaging formats. Furthermore, it is possible that the use of emerging adulthood as a developmental framework and the inclusion of students as part of the project team across all stages of resource development and evaluation contributed to the creation of materials that were particularly relevant for students and were ultimately very well received.

A small proportion of students (35/242, 14.50%) reported never using the digital resource and strategies over the duration of the study. While issues with resource uptake and use were expected, given earlier research findings [36,38,51], it was encouraging that the majority of participants (77/242, 83.49%) reported at least some use of the self-directed website in this study. Exclusion of the subgroup of participants reporting no uptake did not impact the findings of acceptability, revealing comparably high levels of satisfaction across both study groups over time.

In terms of effectiveness, stress and engagement in unhealthy coping behaviors both decreased in the directed and nondirected groups, with no changes observed in the comparison group. Overall, these findings suggest that using the digital resource led to improvements in stress and unhealthy coping; however, there was no added benefit of the screening and referral approach. It is possible that screening had no impact in this study because (1) the researcher-developed measure may not have been sensitive enough to identify groups of need that were meaningfully distinct, or (2) students' need for support was limited in variability in the study sample. If the sample included a greater proportion of students demonstrating a high need for stress management and healthy coping support, they may have benefited to a greater extent from receiving personalized resources.

Finally, the 2 resource groups (directed and nondirected) were merged to examine the effectiveness of the overall digital resource against the comparison group for the same outcomes (ie, stress, coping self-efficacy, coping behaviors, and well-being). Findings revealed significant improvements across stress and coping, although there was no effect on well-being. As hypothesized, stress and unhealthy coping decreased, whereas coping self-efficacy and healthy coping increased from baseline to follow-up among the resource group, with no changes detected in the comparison group. Additionally, the pattern of change was similar across the outcomes where changes were detected for stress and healthy coping between baseline and post timepoints, and changes for coping self-efficacy and unhealthy coping detected between post and follow-up timepoints. Contrary to what was expected, there were no changes in well-being across any of the groups over time. This finding contradicts that of Chung et al [36], who reported significant improvements in well-being (using the same measure) following students' use of a digital self-directed mindfulness program for university students. However, the timeline between baseline and follow-up assessments was shorter in this study (10 wk) in comparison to the 14 weeks between baseline and follow-up in the study by Chung et al [36]. It is therefore possible that additional time is needed to detect changes in subjective well-being in response to engagement with self-directed programming.

Taken together, the findings support the effectiveness of sharing stress-management and healthy coping resources on a self-directed digital platform for improving university students' stress and coping outcomes while demonstrating that the web-based resource was well-received. This study builds on the emerging evidence base highlighting the promise of enhancing university student stress management and coping capacity through universal, digital, self-directed supports [36,69]. Furthermore, findings demonstrate the potential value of extending low-intensity support options (ie, lowest steps within Stepped Care 2.0) [52] beyond the context of clinical service delivery to benefit students [70]. Given problems with help-seeking on campus [43,44], the integration of low-intensity, self-directed stress-management and coping support across the whole university can function to proactively connect students with evidence-based resources.

Contributions

The unique contribution of this study towards research and practice in supporting university students' stress management and healthy coping is threefold. First, this study contributes to the small but growing evidence base demonstrating the feasibility, acceptability, and effectiveness of low-resource, self-directed programming for supporting students' stress and coping outcomes in demanding university environments [36,69]. Second, this study responds to calls for enhancing access to freely available and trustworthy digital resources for managing stress and coping capacity as a supplement to existing mental health services on campus [39,42,71]. Similarly, this study responds to calls to specifically promote the availability of evidence-based strategies for healthy coping in university environments to support coping capacity and mitigate the negative impacts of engaging in unhealthy coping behaviors

[7,22,41]. Third, this study presented the first adaptation of the clinical screening and referral to stepped care approach for use across the general university student population to connect them with lower-intensity resources proportional to their reported level of need for stress-management and healthy coping information. While there was no evidence for a differential benefit of this adapted approach in this study, the results suggest that the screening and directing approach may vary in its effectiveness if used with those with low needs and may only be beneficial when targeting those with a more severe need for support around stress and coping. Last, the web-based, self-directed resource format tested in this study is scalable to other higher education contexts and adaptable to university student populations. The current format allows for low-resource, wide-reaching, and sustainable implementation of student stress and coping support compared with more resource-intensive formats, such as in-person or synchronous options. While the format is inherently scalable, challenges exist at the development and implementation stages, including the initial investment in material development, integration with institutional digital infrastructure, and promotion to ensure student use and engagement. Nonetheless, this upfront effort is worthwhile, as the resource offers sustainable, flexible support with demonstrated benefits for university student stress and coping.

Limitations and Future Directions

Study findings must be interpreted with consideration of the following limitations. First, the timeline of the evaluation study was constrained to a relatively brief 10-week period. Although this timeframe allowed for a focused examination of the specific variables under consideration, it also limits the ability to capture longer-term effects or variations that could emerge over an extended period. Future studies with extended timelines are warranted to explore the sustainability and long-term impacts of web-based, self-directed resources to support university student stress management and coping capacity. Second, one of the measures used (ie, the Coping Index; CI) [64], exhibited poor internal consistency within the health and unhealthy coping subscales. While it was included in this study, given the measure's direct alignment with the theoretical foundations of the study (ie, health theory of coping) [7], caution is advised for future uses of this measure in research in the absence of psychometric validation. Third, students identifying as women were overrepresented in the study sample, which impacts the generalizability of findings. While this is commonly observed across social science research [72], it is crucial for future studies to explore means of engaging participants who represent a more diverse range of gender identities. Fourth, the lack of impact of the screening and directing approach tested within this study could be due to the use of a researcher-developed screening questionnaire and algorithm to facilitate the directing. It is possible that the screening questionnaire was not effective in delineating low, moderate, or high need groups. Future research could consider establishing the validity and sensitivity of the screener measure ahead of examining the effectiveness of the screening and directing approach in the context of an intervention. Fifth, although the web-based resource tested in this study was designed to enhance access to stress and coping supports in the university context, it should be acknowledged that access to reliable internet, personal devices, or private spaces to engage with the content is not universal. These issues present barriers and may affect the generalizability of the findings and the scalability of the resource across diverse higher education settings. As research on digital mental health and well-being programming advances, it is essential to consider and address barriers related to digital equity to ensure broad and inclusive access to support. Finally, a notable limitation in this study is the absence of consideration of intraindividual identity factors (eg, gender, racial/ethnic identity) or lived experience (eg, history of mental illness and/or trauma). Although this study demonstrates the acceptability and effectiveness of a web-based, self-directed resource for supporting university students' stress-management and coping capacity, what remains to be explored is the potentially differential acceptability and effectiveness of the self-directed support option as a function of intraindividual identity factors.

Conclusions

In summary, this study highlights the acceptability and effectiveness of a self-directed, web-based resource providing evidence-based stress-management and healthy coping strategies for university students. Results indicate that students tended to like the overall resource and were satisfied with the content and format of the information presented, although there was no added benefit of the screening and directing approach in this study. Importantly, students' engagement with the resource and use of the strategies led to improvements in stress, their belief in their capacity to cope, and their engagement in healthier coping behaviors. Thus, the web-based resource evaluated in this study demonstrates promise for supplementing existing mental health services on campus to provide additional support for managing stress and enhancing coping capacity among university students.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to confidentiality agreements but are available from the corresponding author on reasonable request.

Authors' Contributions

BNB made substantial contributions to the study's conceptualization, methodological design, development of the resource tested in this study, data collection, data analysis, interpretation of the results, manuscript writing, and editing of the final manuscript. JM and NLH made substantial contributions to the study's conceptualization, design, data analysis, interpretation of the results, critical revision of the manuscript drafts, and approved the final version. LB and SC made substantial contributions to the development of the resource tested in this study, critical revision of manuscript drafts, and editing of the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental file providing additional details on participants' ratings of resource acceptability, the screening measure and algorithm, and screenshots of the web-based resource tested in this study.
[PDF File, 496 KB - mental_v13i1e74205_app1.pdf]

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 440 KB - mental_v13i1e74205_app2.pdf]

References

1. American College Health Association. National college health assessment III undergraduate student reference group executive summary. : Silver Spring; 2022 URL: https://www.acha.org/wp-content/uploads/2024/07/NCHA-III_SPRING_2022_UNDERGRAD_REFERENCE_GROUP_EXECUTIVE_SUMMARY.pdf [accessed 2024-12-01]
2. American College Health Association. National college health assessment III Canadian reference group executive summary. : Silver Spring; 2022 URL: https://www.acha.org/wp-content/uploads/2024/07/NCHA-III_SPRING_2022_CANADIAN_REFERENCE_GROUP_EXECUTIVE_SUMMARY.pdf [accessed 2024-12-01]
3. Sharp J, Theiler S. A review of psychological distress among university students: pervasiveness, implications and potential points of intervention. Int J Adv Counselling 2018 Sep;40(3):193-212. [doi: 10.1007/s10447-018-9321-7]
4. Stallman HM. Psychological distress in university students: a comparison with general population data. Aust Psychol 2010 Dec 1;45(4):249-257. [doi: 10.1080/00050067.2010.482109]
5. Harrer M, Adam SH, Fleischmann RJ, et al. Effectiveness of an internet- and app-based intervention for college students with elevated stress: randomized controlled trial. J Med Internet Res 2018 Apr 23;20(4):e136. [doi: 10.2196/jmir.9293] [Medline: 29685870]
6. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869. [doi: 10.2196/12869] [Medline: 31333198]
7. Stallman HM. Health theory of coping. Aust Psychol 2020 Aug 1;55(4):295-306. [doi: 10.1111/ap.12465]
8. Brown JSL. Student mental health: some answers and more questions. J Ment Health 2018 Jun;27(3):193-196. [doi: 10.1080/09638237.2018.1470319] [Medline: 29768071]
9. Hill M, Farrelly N, Clarke C, Cannon M. Student mental health and well-being: overview and future directions. Ir J Psychol Med 2024 Jun;41(2):259-266. [doi: 10.1017/ipm.2020.110] [Medline: 33243317]
10. Van de Velde S, Buffel V, Bracke P, et al. The COVID-19 international student well-being study. Scand J Public Health 2021 Feb;49(1):114-122. [doi: 10.1177/1403494820981186] [Medline: 33406995]
11. Conley CS, Kirsch AC, Dickson DA, Bryant FB. Negotiating the transition to college: developmental trajectories and gender differences in psychological functioning, cognitive-affective strategies, and social well-being. Emerg Adulthood 2014;2(3):195-210. [doi: 10.1177/2167696814521808]
12. Arnett JJ. Emerging adulthood. a theory of development from the late teens through the twenties. Am Psychol 2000 May;55(5):469-480. [doi: 10.1037/0003-066X.55.5.469] [Medline: 10842426]
13. Arnett JJ. Emerging Adulthood: The Winding Road from the Late Teens Through the Twenties: Oxford University Press; 2004. [doi: 10.1093/oxfordhb/9780199795574.013.9]
14. Swanson JA. Trends in literature about emerging adulthood: review of empirical studies. Emerg Adulthood 2016;4(6):391-402. [doi: 10.1177/2167696816630468]
15. Syed M, Mitchell LL. Race, ethnicity, and emerging adulthood: retrospect and prospects. Emerg Adulthood 2013;1(2):83-95. [doi: 10.1177/2167696813480503]
16. Böke BN, Mills DJ, Mettler J, Heath NL. Stress and coping patterns of university students. J Coll Stud Dev 2019;60(1):85-103. [doi: 10.1353/csd.2019.0005]
17. Bukobza G. Relations between rebelliousness, risk-taking behavior, and identity status during emerging adulthood. Identity (Mahwah, N J) 2009 Apr 21;9(2):159-177. [doi: 10.1080/15283480802676932]
18. Sussman S, Arnett JJ. Emerging adulthood: developmental period facilitative of the addictions. Eval Health Prof 2014;37(2):147-155. [doi: 10.1177/0163278714521812]

19. Slimmen S, Timmermans O, Mikolajczak-Degrauwe K, Oenema A. How stress-related factors affect mental wellbeing of university students a cross-sectional study to explore the associations between stressors, perceived stress, and mental wellbeing. PLoS One 2022;17(11):e0275925. [doi: 10.1371/journal.pone.0275925] [Medline: 36342914]
20. Cunningham S, Duffy A. Investing in our future: importance of postsecondary student mental health research. Can J Psychiatry 2019 Feb;64(2):79-81. [doi: 10.1177/0706743718819491]
21. Munthali RJ, Richardson CG, Pei J, Westenberg JN, Munro L, Auerbach RP, et al. Patterns of anxiety, depression, and substance use risk behaviors among university students in Canada. J Am Coll Health 2023:1-11. [doi: 10.1080/07448481.2023.2277201]
22. Stallman HM, Lipson SK, Zhou S, Eisenberg D. How do university students cope? an exploration of the health theory of coping in a US sample. J Am Coll Health 2022;70(4):1179-1185. [doi: 10.1080/07448481.2020.1789149]
23. Amanvermez Y, Rahmadiana M, Karyotaki E, et al. Stress management interventions for college students: a systematic review and meta-analysis. Clinical Psychology: Science and Practice 2020;30(4):423-444. [doi: 10.1111/cpsp.12342]
24. Ma L, Zhang Y, Cui Z. Mindfulness-based interventions for prevention of depressive symptoms in university students: a meta-analytic review. Mindfulness (N Y) 2019 Nov;10(11):2209-2224. [doi: 10.1007/s12671-019-01192-w]
25. Ang WHD, Lau ST, Cheng LJ, et al. Effectiveness of resilience interventions for higher education students: a meta-analysis and metaregression. J Educ Psychol 2022;114(7):1670-1694. [doi: 10.1037/edu0000719]
26. Worsley JD, Pennington A, Corcoran R. Supporting mental health and wellbeing of university and college students: a systematic review of review-level evidence of interventions. PLoS One 2022;17(7):e0266725. [doi: 10.1371/journal.pone.0266725] [Medline: 35905058]
27. Conley CS, Durlak JA, Kirsch AC. A meta-analysis of universal mental health prevention programs for higher education students. Prev Sci 2015 May;16(4):487-507. [doi: 10.1007/s11121-015-0543-1] [Medline: 25744536]
28. Harrer M, Adam SH, Baumeister H, et al. Internet interventions for mental health in university students: a systematic review and meta-analysis. Int J Methods Psychiatr Res 2019 Jun;28(2):e1759. [doi: 10.1002/mpr.1759] [Medline: 30585363]
29. Bolinski F, Boumparis N, Kleiboer A, Cuijpers P, Ebert DD, Riper H. The effect of e-mental health interventions on academic performance in university and college students: a meta-analysis of randomized controlled trials. Internet Interv 2020 Apr;20:100321. [doi: 10.1016/j.invent.2020.100321] [Medline: 32382515]
30. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 6;20(6):e199. [doi: 10.2196/jmir.9275] [Medline: 29875089]
31. Gabrielli S, Rizzi S, Bassi G, et al. Engagement and effectiveness of a healthy-coping intervention via chatbot for university students during the COVID-19 pandemic: mixed methods proof-of-concept study. JMIR mHealth uHealth 2021 May 28;9(5):e27965. [doi: 10.2196/27965] [Medline: 33950849]
32. Lattie EG, Lipson SK, Eisenberg D. Technology and college student mental health: challenges and opportunities. Front Psychiatry 2019;10:246. [doi: 10.3389/fpsyt.2019.00246] [Medline: 31037061]
33. Fischer R, Bortolini T, Karl JA, Zilberberg M, Robinson K, Rabelo A, et al. Rapid review and meta-meta-analysis of self-guided interventions to address anxiety, depression, and stress during COVID-19 social distancing. Front Psychol 2020;11:563876. [doi: 10.3389/fpsyg.2020.5638076]
34. Cuijpers P, Donker T, Johansson R, Mohr DC, van Straten A, Andersson G. Self-guided psychological treatment for depressive symptoms: a meta-analysis. PLoS One 2011;6(6):e21274. [doi: 10.1371/journal.pone.0021274] [Medline: 21712998]
35. Karyotaki E, Riper H, Twisk J, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms. JAMA Psychiatry 2017 Apr 1;74(4):351. [doi: 10.1001/jamapsychiatry.2017.0044]
36. Chung J, Mundy ME, McKenzie S. A self-managed online mindfulness program in a university-wide learning management system orientation site: a real-world ecological validation study. Front Psychol 2022;13:869765. [doi: 10.3389/fpsyg.2022.869765] [Medline: 35602693]
37. Fleischmann RJ, Harrer M, Zarski AC, Baumeister H, Lehr D, Ebert DD. Patients’ experiences in a guided internet- and app-based stress intervention for college students: a qualitative study. Internet Interv 2018 Jun;12:130-140. [doi: 10.1016/j.invent.2017.12.001] [Medline: 30135777]
38. Lillevoll KR, Vangberg HCB, Griffiths KM, Waterloo K, Eisemann MR. Uptake and adherence of a self-directed internet-based mental health intervention with tailored e-mail reminders in senior high schools in Norway. BMC Psychiatry 2014 Jan 21;14:1-11. [doi: 10.1186/1471-244X-14-14] [Medline: 24443820]
39. Becker TD, Torous JB. Recent developments in digital mental health interventions for college and university students. Curr Treat Options Psych 2019 Sep;6(3):210-220. [doi: 10.1007/s40501-019-00178-8]
40. Oti O, Pitt I. Online mental health interventions designed for students in higher education: a user-centered perspective. Internet Interv 2021 Dec;26:100468. [doi: 10.1016/j.invent.2021.100468] [Medline: 34703772]
41. Reis A, Saheb R, Parish P, Earl A, Klupp N, Sperandei S. How I cope at university: self‐directed stress management strategies of Australian students. Stress and Health 2021;37(5):1010-1025. [doi: 10.1002/smi.3058]

42. Ahuvia IL, Sung JY, Dobias ML, Nelson BD, Richmond LL, London B, et al. College student interest in teletherapy and self-guided mental health supports during the COVID-19 pandemic. J Am Coll Health 2022:1-7. [doi: 10.1080/07448481.2022.2062245]
43. Bourdon JL, Moore AA, Long EC, Kendler KS, Dick DM. The relationship between on-campus service utilization and common mental health concerns in undergraduate college students. Psychol Serv 2020;17(1):118-126. [doi: 10.1037/ser0000296]
44. Dunley P, Papadopoulos A. Why is it so hard to get help? Barriers to help-seeking in postsecondary students struggling with mental health issues: a scoping review. Int J Ment Health Addiction 2019 Jun;17(3):699-715. [doi: 10.1007/s11469-018-0029-z]
45. Lakhtakia T, Torous J. Current directions in digital interventions for mood and anxiety disorders. Curr Opin Psychiatry 2022;35(2):130-135. [doi: 10.1097/YCO.0000000000000772]
46. Eisenberg D, Hunt J, Speer N, Zivin K. Mental health service utilization among college students in the United States. J Nerv Ment Dis 2011 May;199(5):301-308. [doi: 10.1097/NMD.0b013e3182175123] [Medline: 21543948]
47. Cho S, Bastien L, Petrovic J, Böke BN, Heath NL. The role of mental health stigma in university students' satisfaction with web-based stress management resources: intervention study. JMIR Form Res 2024 Apr 4;8(1):e50018. [doi: 10.2196/50018] [Medline: 38573758]
48. Böke BN, Joly M, Bastien L, Heath NL. Keep it brief: can a 4-item stress screener predict university adjustment over 18 months? High Educ Res Dev 2024 Jul 3;43(5):1026-1039. [doi: 10.1080/07294360.2023.2291061]
49. Hasking PA, Robinson K, McEvoy P, et al. Development and evaluation of a predictive algorithm and telehealth intervention to reduce suicidal behavior among university students. Psychol Med 2024 Apr;54(5):971-979. [doi: 10.1017/S0033291723002714] [Medline: 37732419]
50. King CA, Eisenberg D, Zheng K, et al. Online suicide risk screening and intervention with college students: a pilot randomized controlled trial. J Consult Clin Psychol 2015 Jun;83(3):630-636. [doi: 10.1037/a0038805] [Medline: 25688811]
51. Rith-Najarian LR, Boustani MM, Chorpita BF. A systematic review of prevention programs targeting depression, anxiety, and stress in university students. J Affect Disord 2019 Oct 1;257:568-584. [doi: 10.1016/j.jad.2019.06.035] [Medline: 31326690]
52. Cornish P. Stepped Care 2.0: A Paradigm Shift in Mental Health: Springer Nature; 2020. [doi: 10.1007/978-3-030-48055-4]
53. Cornish PA, Berry G, Benton S, et al. Meeting the mental health needs of today's college student: reinventing services through Stepped Care 2.0. Psychol Serv 2017 Nov;14(4):428-442. [doi: 10.1037/ser0000158] [Medline: 29120201]
54. Web content accessibility guidelines (WCAG). World Wide Web Consortium. 2008 Dec 11. URL: https://www.w3.org/TR/WCAG20/ [accessed 2022-10-20]
55. Cohen S, Williamson GM. Perceived stress in a probability sample of the United States. In: Spacapan S, Oskamp S, editors. The Social Psychology of Health: Sage; 1988:31-67 URL: https://www.cmu.edu/dietrich/psychology/stress-immunity-disease-lab/scales/pdf/cohen,-s.--williamson,-g.-1988.pdf [accessed 2026-01-02]
56. Chesney MA, Neilands TB, Chambers DB, Taylor JM, Folkman S. A validity and reliability study of the coping self-efficacy scale. Br J Health Psychol 2006 Sep;11(Pt 3):421-437. [doi: 10.1348/135910705X53155] [Medline: 16870053]
57. Russell D, Peplau LA, Cutrona CE. The revised UCLA loneliness scale: concurrent and discriminant validity evidence. J Pers Soc Psychol 1980;39(3):472-480. [doi: 10.1037/0022-3514.39.3.472]
58. Zimet GD, Dahlem NW, Zimet SG, Farley GK. The multidimensional scale of perceived social support. J Pers Assess 1988 Mar;52(1):30-41. [doi: 10.1207/s15327752jpa5201_2]
59. Armstrong S, Oomen-Early J. Social connectedness, self-esteem, and depression symptomatology among collegiate athletes versus nonathletes. J Am Coll Health 2009;57(5):521-526. [doi: 10.3200/JACH.57.5.521-526] [Medline: 19254893]
60. Kirkpatrick JD, Kirkpatrick WK. Kirkpatrick's Four Levels of Training Evaluation: Association for Talent Development; 2016.
61. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983;24(4):385-396. [doi: 10.2307/2136404]
62. Denovan A, Dagnall N, Dhingra K, Grogan S. Evaluating the perceived stress scale among UK university students: implications for stress measurement and management. Stud High Educ 2019;44(1):120-133. [doi: 10.1080/03075079.2017.1340445]
63. Luberto CM, Cotton S, McLeish AC, Mingione CJ, O'Bryan EM. Mindfulness skills and emotion regulation: the mediating role of coping self-efficacy. Mindfulness (N Y) 2014 Aug;5(4):373-380. [doi: 10.1007/s12671-012-0190-6]
64. Stallman HM, Beaudequin D, Hermens DF, Eisenberg D. Modelling the relationship between healthy and unhealthy coping strategies to understand overwhelming distress: A Bayesian network approach. Journal of Affective Disorders Reports 2021 Jan;3. [doi: 10.1016/j.jadr.2020.100054]
65. Stallman HM. Efficacy of the My Coping Plan mobile application in reducing distress: a randomised controlled trial. Clin Psychol (Aust Psychol Soc) 2019 Nov 1;23(3):206-212. [doi: 10.1111/cp.12185]
66. Meyers LS, Gamst G, Guarino AJ. Applied Multivariate Research: Design and Interpretation, 3rd edition: Sage Publications Ltd; 2017. [doi: 10.4135/9781071802687]

67. Tennant R, Hiller L, Fishwick R, et al. The Warwick-Edinburgh mental well-being scale (WEMWBS): development and UK validation. Health Qual Life Outcomes 2007 Nov 27;5(1):1-13. [doi: 10.1186/1477-7525-5-63] [Medline: 18042300]
68. Neal DM, Campbell AJ, Williams LY, Liu Y, Nussbaumer D. "I did not realize so many options are available": cognitive authority, emerging adults, and e-mental health. Lib Inf Sci Res 2011 Jan;33(1):25-33. [doi: 10.1016/j.lisr.2010.07.015]
69. Bastien L, Boke BN, Mettler J, et al. Peer-presented versus mental health service provider-presented mental health outreach programs for university students: randomized controlled trial. JMIR Ment Health 2022 Jul 22;9(7):e34168. [doi: 10.2196/34168] [Medline: 35762935]
70. Ryan ML, Shochet IM, Stallman HM. Universal online interventions might engage psychologically distressed university students who are unlikely to seek formal help. Adv Ment Health 2010 Aug;9(1):73-83. [doi: 10.5172/jamh.9.1.73]
71. Montagni I, Tzourio C, Cousin T, Sagara JA, Bada-Alonzi J, Horgan A. Mental health-related digital use by university students: a systematic review. Telemed J E Health 2020 Feb;26(2):131-146. [doi: 10.1089/tmj.2018.0316] [Medline: 30888256]
72. Becker R. Gender and survey participation: an event history analysis of the gender effects of survey participation in a probability-based multi-wave panel study with a sequential mixed-mode design. Methods Data Anal 2022;16(1):30. [doi: 10.12758/mda.2021.08]

Abbreviations

CI: Coping Index
CSE: Coping Self-Efficacy
MET: main effect of time
PSS: Perceived Stress Scale
SC2.0: Stepped Care 2.0
T1: baseline
T2: post
T3: follow-up
WEMWBS: Warwick-Edinburgh Mental Well-Being Scale

Edited by J Torous; submitted 21.Mar.2025; peer-reviewed by G Jones, S Milner; revised version received 28.Aug.2025; accepted 04.Sep.2025; published 23.Jan.2026.

Please cite as:
Böke BN, Mettler J, Bastien L, Cho S, Heath N
Examining the Acceptability and Effectiveness of a Self-Directed, Web-Based Resource for Stress and Coping in University: Randomized Controlled Trial
JMIR Ment Health 2026;13:e74205
URL: https://mental.jmir.org/2026/1/e74205
doi:10.2196/74205

Telebehavioral Health, In-Person, and Hybrid Modalities of Treatment Delivery Among US Service Members: Longitudinal Observational Study

Kristen H Walter[1], PhD; Lisa H Glassman[1,2], PhD; Jordan A Levine[1,2], MPH; Hee-Jin Jun[3], MPH, SCD; James F Bonkowski[1,2], BBA; Samuel Y Chung[1,2], PhD; Emily A Schmied[3], MPH, PhD

[1]Naval Health Research Center, 140 Sylvester Road, San Diego, CA, United States

[2]Leidos, Inc, Naval Health Research Center, San Diego, CA, United States

[3]School of Public Health, San Diego State University, San Diego, CA, United States

Corresponding Author:
Kristen H Walter, PhD
Naval Health Research Center, 140 Sylvester Road, San Diego, CA, United States

Abstract

Background: The availability of telebehavioral health care dramatically increased in response to the COVID-19 pandemic among both civilian and military populations. After the restrictions were lifted, telebehavioral health use decreased but remained elevated compared to before the pandemic. Examining the use of treatment modalities and how they relate to care metrics can inform the future delivery of behavioral health care.

Objective: This study aimed to explore behavioral health use patterns by treatment modality (telehealth, in-person, and hybrid care) among active duty service members with at least 1 of 12 behavioral health conditions. Treatment modality groups were also compared on the number of visits and between-visit intervals to determine the association with care metrics.

Methods: The study included 588,928 active duty service members who completed at least 6 months of continuous service during the study period (October 1, 2015, to September 30, 2021) and received care for at least 1 behavioral health condition of interest. Personnel and demographic data were matched with medical reimbursement records. Diagnostic and treatment procedure codes were extracted for each health care visit. For each service member in the study population, the total number of behavioral health visits, modality of each visit, and average duration of time between visits were calculated.

Results: Overall, 59.57% (350,843/588,928) of service members received only in-person care during the 6-year study period, 4.12% (24,245/588,928) received only telehealth, and 36.31% (213,840/588,928) received hybrid care. For 8 (66.7%) of the 12 behavioral health conditions (eg, alcohol use disorder, attention-deficit/hyperactivity disorder, generalized anxiety disorder, major depressive disorder, panic disorder, posttraumatic stress disorder, substance use disorder, and suicidal behavior), service members were more likely to receive hybrid care, whereas the other 4 (41.7%) conditions (eg, acute stress disorder, adjustment disorder, insomnia, and suicidal ideation) were more likely to be associated with in-person care. Service members who received hybrid care averaged 8 times more visits than those using only telehealth and 3 times more visits than those receiving only in-person care. For most conditions, service members who received in-person care only averaged the longest intervals between visits, whereas those who used telehealth care only averaged the shortest intervals. Among specific behavioral health conditions, average intervals were longest among those with attention-deficit/hyperactivity disorder, acute stress disorder, and insomnia (79 - 89 d) and shortest among those with suicidal behavior, substance use disorder, and alcohol use disorder (25 - 38 d).

Conclusions: Telebehavioral health care was commonly used in combination with in-person care and associated with more health care visits and the least amount of time between visits, revealing advantages of offering telehealth within the Military Health System. Findings support a flexible care delivery approach that includes various modalities, such as telehealth, in-person, and hybrid options to address the behavioral health needs of service members.

(JMIR Ment Health 2026;13:e83809) doi:10.2196/83809

KEYWORDS

telemedicine; teletherapy; mHealth; mobile health; eHealth; digital health; military; mental health; psychotherapy; therapy

Introduction

Background

Behavioral health care delivered via telehealth is not a new practice in the United States; however, the COVID-19 pandemic led to a rapid increase in the capability and accessibility of this treatment modality among both civilian and military populations. Telehealth is broadly defined by the American Telemedicine Association as a "mode of delivering healthcare services through the use of telecommunications technologies ...by a healthcare practitioner to a patient at a different physical location than the healthcare practitioner" [1], and the term has been used interchangeably with telemedicine, teletherapy, mobile health, eHealth, and digital health delivery across the literature. Telebehavioral health served a critical need during the pandemic, especially as nationally representative civilian data [2] showed dramatic increases in behavioral health symptoms and distress resulting from the pandemic. Among military populations across 5 countries, resiliency was demonstrated early in the pandemic. However, mental health worsened, and stress levels increased over time for certain subpopulations, such as service members who were deployed to provide aid and assistance in response to COVID-19 [3]. Telehealth provided a way to manage behavioral health needs across civilian and military populations during the pandemic by maintaining social distancing requirements and addressing gaps in care delivery. These characteristics support the ongoing and increased availability of this treatment modality.

For telehealth during the COVID-19 pandemic more generally, nationally representative civilian data showed that rates of telehealth visits increased by 17% during the first 6 calendar months of 2020 (from 0.8 to 17.8 visits per 1000 enrollees), whereas in-person visits decreased 26% (from 102.7 to 76.3 per 1000 enrollees) [4]. The increase in telehealth use was also evident among active duty service members, as telehealth visits increased by 20-fold, with 2,891,865 visits in 2020 compared to 138,138 in 2019 [5]. For behavioral health conditions more specifically, telehealth visits were approximately 25% higher during March through September 2020 compared with the same period in 2019 [6]. Rates of telebehavioral health visits among active duty service members peaked in April 2020 and declined by mid-June [6-8], a trend also observed across 11 behavioral health conditions [7]. Clark et al [6] further observed that military telebehavioral health visit rates stabilized after June 2020 but were consistently elevated compared with the prior year, implying that telehealth had become a larger part of the health care landscape and an option more readily available than before the pandemic. However, less is known about how the use of telebehavioral health fits within service members' overall treatment use.

Although telehealth rates increased dramatically at the start of the pandemic, telehealth use had steadily increased over the prior decade [9,10]. Telehealth offers solutions to several barriers to in-person behavioral health care for patients, providers, and health care systems. For example, across patient populations, telebehavioral health can eliminate geographic constraints by delivering care to patients in remote and rural areas and those with provider shortages [11-15]. In the military, service members in austere, far-forward, and shipboard locations can receive telebehavioral health care, potentially reducing resource- and cost-intensive medical evacuations [16], for which behavioral health conditions are a leading cause [17,18]. Telebehavioral health also accommodates patients who prefer to receive care in their own home due to concerns related to mobility or to privacy and/or psychological comfort [11,12,19]. A private setting of the patient's choosing may facilitate care seeking among those who otherwise would not receive care in a traditional medical facility due to perceived stigma—a barrier for many populations, especially the military [12,20-25]. These potential advantages of telehealth may also vary within military contexts. For example, an officer receiving behavioral health care on base may benefit from the privacy of telehealth, whereas a junior-enlisted service member living in barracks or quarters may have difficulty finding a private space to attend telehealth sessions. The options of telehealth and other delivery modalities allow service members the opportunity to access behavioral health care in ways that may best address their treatment needs.

Importantly, evidence suggests that psychotherapies delivered via telehealth are generally as effective at reducing behavioral health symptoms as in-person treatment in both civilian and military populations [22,26-28]. Furthermore, many patients receiving telebehavioral health reported similar relationship building with their therapist, comparable to in-person treatment [29,30]. For providers, telebehavioral health can result in increased clinical efficiency by reducing time to care initiation [31], shortening care episodes [32], and facilitating faster appointments and decreasing the frequency of no-shows [33]. Both patients and providers may experience a reduction in treatment-related expenses due to lower costs associated with travel, transportation, time, and missed work [19,34-37]. Health systems also benefit through reduced use of medical supplies, lower facility fees, and lower overhead costs [38]. In the Military Health System (MHS), the opportunity cost savings of telebehavioral health were determined to be over US $1.1 million for officers and US $740,000 for enlisted service members compared with in-person visits in 2020 [31]. Taken together, telebehavioral health provides numerous advantages for patients, providers, and health care systems and overcomes barriers associated with distance, preference, and cost.

Although telebehavioral health offers many benefits for care delivery, there are challenges, and it may not be suited for all patients. A main concern involves technology, such as competence with technology, comfort communicating over video conferencing, internet quality and connectivity, experience with software, and hardware that can restrict optimal performance [11,14,15,24,36,37,39]. Technology access issues may also exist in the very military settings where there is a critical need for telehealth care—austere, far forward, and shipboard locations. Socioeconomic factors affecting technological access and familiarity could also create a situation of *digital exclusion* from telehealth [40,41], leading to health care disparities. For example, while telehealth can promote patient privacy and comfort, some may lack a safe or private space, experience disruptions in their environment [11,14,42], or may not be able to afford fast connection speeds or updated devices that support telehealth use [43]. Additionally, although research generally shows comparable outcomes for telehealth and in-person psychotherapy [22,26-28], there are clinical subgroups that appear to benefit more from in-person treatment. Specifically, treatments for depression showed better outcomes when delivered in person than through telehealth [44]. Other subgroups with higher hopelessness or anxiety symptom severity also showed better outcomes for in-person care versus telehealth [26], and individuals with greater symptom severity and behavioral health comorbidities may be better matched to in-person treatment [45,46]. Given the differences between telebehavioral health and in-person treatment, it is important to understand the use of these modalities over time, including when used in combination.

Objectives

Most existing research examining the expansion of telehealth use in response to the pandemic either assessed telehealth alone or changes in telehealth and in-person care as single modalities [4-6,32]; limited research has explored the combined use of telehealth and in-person care (ie, "hybrid" care). An exception is a recent RAND study conducted by Hepner et al [8], who reported that most service members who began behavioral health treatment in the early months of the pandemic received a hybrid of telehealth and in-person visits. This study examined the corresponding 6-month periods in 2019 and 2020 and focused on 3 diagnoses (ie, posttraumatic stress disorder [PTSD], depression, and substance use disorder [SUD]).

This study built on these findings by evaluating behavioral health use patterns by treatment modality (ie, telehealth, in-person, and hybrid care) among active duty service members with at least one of the 12 behavioral health diagnoses of interest over a 6-year period that extended to September 2021. Furthermore, the modalities were compared on the number of visits and between-visit intervals to determine whether the modalities were associated with care metrics. Study results can inform the future delivery of behavioral health care to service members, supporting the aims of the Department of Defense [47] and the Defense Health Agency [48]. On a broader level, this study raises considerations for flexible delivery, personal choice, and shared decision-making in behavioral health care [22,49,50].

Methods

Data Sources

The base population consisted of active duty military service members with at least 6 months of continuous service during the study period (October 1, 2015, to September 30, 2021) and who received care for at least 1 behavioral health condition of interest. The study period began in 2015, as telehealth care was seldom used in the MHS before this time [7,51,52]. The 12 behavioral health conditions of interest included acute stress disorder (ASD), adjustment disorder, alcohol use disorder (AUD), attention-deficit/hyperactivity disorder (ADHD), generalized anxiety disorder (GAD), insomnia, major depressive disorder (MDD), suicidal behavior, panic disorder, PTSD, SUD, and suicidal ideation. These selected behavioral health conditions are common in military populations and have been explored in prior research using similar data sources [7,53].

Personnel and demographic data were derived from the Career History Archival Medical and Personnel System and then matched with specific diagnoses from medical reimbursement records housed in the MHS Data Repository (MDR). The MDR contains health care data from TRICARE (ie, the military health care program) or TRICARE-reimbursed facilities, which include both military and civilian treatment facilities. The medical data captured reflects services used that were reimbursed by TRICARE, whether elective or mandated by a service member's command (eg, command-directed substance use treatment). Therefore, MDR data represent the *use* of health care services, but not necessarily a *preference* for the care received.

Behavioral health diagnoses were identified based on records containing both (1) *International Classification of Diseases, 10th Revision* (ICD-10) [54], codes denoting conditions of interest; and (2) corresponding Current Procedural Terminology or Healthcare Common Procedure Coding System codes indicating the treatment modality (telehealth vs in-person) of each health care visit. Visits were only included if behavioral health treatment was provided for at least one of the eligible diagnoses. Behavioral health treatment included services such as individual psychotherapy, family or group therapy, diagnostic or psychological testing, health behavior interventions, psychiatry evaluation and management, and substance use treatment and intervention. Behavioral health visits were further classified as either in person or telehealth using relevant Current Procedural Terminology and Healthcare Common Procedure Coding System codes [8]. For example, 99443 designates a telephone evaluation or management visit lasting 21 to 30 minutes, and modifier code "95" denotes a synchronous audio-video visit delivered to a patient not located at a military treatment facility. These codes can be found in Multimedia Appendix 1.

For each service member in the sample, the total number of behavioral health visits, the modality of each visit, and the average duration of time between visits were calculated. Demographic data included sex, race and ethnicity, service branch, age, and rank at the first behavioral health visit.

Ethical Considerations

The study protocol was approved by the Naval Health Research Center Institutional Review Board (NHRC.2022.0005) in compliance with all applicable federal regulations. This study used archival data, and therefore, informed consent and compensation were not part of the study. Data were accessed and protected following federal and US Department of Defense regulations.

Statistical Analyses

Descriptive statistics were computed for all demographic variables, behavioral health conditions, and frequency of visits (total number of visits and average time between visits) for the full sample and then separately for each of the 3 care modality groups (ie, telehealth, in-person, and hybrid care). Chi-square tests of independence were used to assess the demographic distribution across treatment delivery modalities. As most demographic characteristics were categorical, post hoc tests were used to identify differences in treatment delivery modality against a reference group within each categorical demographic variable. Reference groups included non-Hispanic White race and ethnicity, the Marine Corps service branch, and junior enlisted rank. Cramer *V* statistics were then computed to determine the effect size of differences (with a large effect size defined as a Cramer *V* value of ≥0.15) between these groups.

Chi-square tests of equal proportions were used to assess statistical significance in the distribution of care across treatment delivery modalities for each of the 12 behavioral health conditions of interest. As the use of telehealth alone was less frequent than only in-person care or the hybrid of telehealth and in-person care during the observation period, post hoc tests were run to determine differences between in-person and hybrid care.

ANOVA tests were computed to assess differences in the average number of visits and the average time between visits across the 3 treatment delivery modality groups, both for care overall and for each specific behavioral health condition. An η^2 statistic assessed the effect size of the results. In cases where the effect size was medium (0.06 - 0.13) or large (≥0.14), a post hoc Tukey test was conducted to identify differences between the 3 treatment delivery modality groups for each behavioral health condition.

Results

Demographics

A total of 622,452 service members received care for the behavioral health conditions of interest during the study period and had at least 6 months of continuous service. After removing those with incomplete personnel records (n=33,524, 5.39%), the study sample consisted of 588,928 service members. Overall, 350,843 (59.57%) service members received only in-person care during the study period, 24,245 (4.12%) received only telehealth, and 213,840 (36.31%) received a hybrid of in-person and telehealth care.

Both omnibus and subsequent post hoc chi-square tests of independence revealed statistically significant differences in care use within each demographic variable (Table 1). However, no calculations produced a large effect size. Analyses between service branches, ranks, and sexes produced small effect sizes (Cramer *V* between 0.04 and 0.09). Specifically, compared with Marines, soldiers and airmen were more likely to use hybrid care, and those in the Coast Guard were more likely to use telehealth care alone. Compared with junior enlisted members, senior enlisted members, officers, and warrant officers were more likely to use telehealth care, both on its own and in conjunction with in-person care. Women were more likely to use a combination of in-person and telehealth services compared with men, who were more likely to use in-person services alone.

Table . Patient demographics by treatment modality.

Characteristics	Overall	Breakdown by treatment modality			Post hoc chi-square tests[a]	
		Both in-person and telehealth	In person only	Telehealth only	*P* value	Cramer *V*
Sex, n (%)						
Female	138,508 (23.60)	56,896 (41.08)	75,998 (54.87)	5614 (4.05)	<.01	0.05
Male	448,335 (76.40)	156,825 (34.98)	272,909 (60.87)	18,601 (4.15)	Reference	—[b]
Unknown	2085	—	—	—	—	—
Race and ethnicity, n (%)						
American Indian–Alaskan Native	16,947 (2.95)	6321 (37.30)	9979 (58.88)	647 (3.82)	<.01	0.01
Asian American–Pacific Islander	25,561 (4.45)	8806 (34.45)	15,593 (61.00)	1162 (4.55)	<.01	0.01
Black–African American	168,480 (29.32)	64,179 (38.09)	97,282 (57.74)	7019 (4.17)	<.01	0.03
Hispanic-Latino	56,647 (9.86)	19,650 (34.69)	34,655 (61.18)	2342 (4.13)	.02	0.01
Multiracial	38,281 (6.66)	15,355 (40.11)	21,374 (55.83)	1552 (4.05)	<.01	0.03
Non-Hispanic White	268,645 (46.76)	94,844 (35.30)	163,066 (60.70)	10,735 (4.00)	Reference	—
Unknown	14,367	—	—	—	—	—
Service branch, n (%)						
Air Force	121,008 (20.61)	46,615 (38.52)	68,601 (56.69)	5792 (4.79)	<.01	0.09
Army	268,136 (45.68)	104,241 (38.88)	154,841 (57.75)	9054 (3.38)	<.01	0.07
Coast Guard	9228 (1.57)	2255 (24.44)	6276 (68.01)	697 (7.55)	<.01	0.06
Marine Corps	64,288 (10.95)	19,223 (29.90)	42,295 (65.79)	2770 (4.31)	Reference	—
Navy	124,366 (21.19)	41,446 (33.33)	77,006 (61.92)	5914 (4.76)	<.01	0.04
Unknown	1902	—	—	—	—	—
Rank, n (%)						
Junior enlisted	250,207 (42.49)	82,039 (32.79)	160,157 (64.01)	8011 (3.20)	Reference	—
Officer or warrant officer	64,066 (10.88)	23,917 (37.33)	35,859 (55.97)	4290 (6.70)	<.01	0.09
Senior enlisted	274,655 (46.64)	107,884 (39.28)	154,827 (56.37)	11,944 (4.35)	<.01	0.08
Age at first visit (y), mean (SD)	28.48 (7.78)	28.89 (7.59)	28.10 (7.84)	30.38 (8.06)	<.01[c]	0.00

[a]These tests analyze the distribution of care modalities against a reference group within each categorical variable (ie, non-Hispanic White race or ethnicity, Marine Corps service branch, and junior enlisted rank).

[b]Not available.

[c]As a continuous variable, age distribution was assessed using a 1-way ANOVA test and corresponding η^2 value.

Care Delivery Modality by Diagnosis

Overall, and irrespective of delivery modality, service members most often received care for adjustment disorder (336,766/588,928, 57%) and insomnia (240,776/588,928, 41%), after which there was a steep drop-off (the next most prevalent condition was AUD 96,509/588,928, 16%; Table 2). Service members infrequently received care for panic disorder (3%), suicidal behavior (0.4%), and suicidal ideation (0.5%).

Table . Patient treatment modality by behavioral health diagnosis.

Diagnosis	Overall, n (%)[a]	Breakdown by treatment modality, n (%)			Post hoc chi-square tests[bcd] (*P* value)
		Both in-person and telehealth	In person only	Telehealth only	
Acute stress disorder	25,304 (4.30)	10,632 (42.02)	13,855 (54.75)	817 (3.23)	<.01
Adjustment disorder	336,766 (57.18)	140,783 (41.80)	186,809 (55.47)	9174 (2.72)	<.01
Alcohol use disorder	96,509 (16.39)	49,339 (51.12)	45,548 (47.20)	1622 (1.68)	<.01
Attention-deficit/hyper-activity disorder	52,337 (8.89)	31,990 (61.12)	19,308 (36.89)	1039 (1.99)	<.01
Generalized anxiety disorder	59,046 (10.03)	34,289 (58.07)	23,158 (39.22)	1599 (2.71)	<.01
Insomnia	240,776 (40.88)	100,936 (41.92)	130,585 (54.24)	9255 (3.84)	<.01
Major depressive disorder	76,641 (13.01)	46,283 (60.39)	28,935 (37.75)	1423 (1.86)	<.01
Panic disorder	15,434 (2.62)	8503 (55.09)	6590 (42.70)	341 (2.21)	<.01
Posttraumatic stress disorder	82,517 (14.01)	49,269 (59.71)	31,639 (38.34)	1609 (1.95)	<.01
Substance use disorder	21,171 (3.76)	12,189 (54.98)	9648 (43.52)	334 (1.51)	<.01
Suicidal behavior	2350 (0.40)	1366 (58.13)	976 (41.53)	8 (0.34)	<.01
Suicidal ideation	2912 (0.49)	1327 (45.57)	1530 (52.54)	55 (1.89)	<.01

[a]The columns add up to a number higher than the total N because many people in the study population had more than one diagnosis.

[b]These post hoc chi-square tests of equal proportion were conducted between in-person care only and combination in-person and telehealth care.

[c]Analyses link behavioral health diagnosis to the visit modality.

[d]The unit of measurement is the patient, not the visit. In this analysis, we are investigating patients' treatment modality overall, rather than the total number of visits administered.

Those with ASD, adjustment disorder, insomnia, and suicidal ideation had among the lowest use of hybrid telehealth and in-person care (42% - 46%), and the highest use of in-person care alone (approximately 55%). Service members with ADHD, GAD, MDD, PTSD, and suicidal behavior were most likely to use a hybrid of in-person and telehealth care (57% - 61%) and least likely to use in-person care alone (37% - 42%). Those with AUD, ADHD, PTSD, SUD, MDD, suicidal ideation, and suicidal behavior were the least likely to use telehealth services alone (0.34% - 2%). All post hoc tests revealed statistically significant differences in the use of in-person alone versus hybrid care for each behavioral health condition ($P<.01$).

Number of Visits

On average, service members attended 10 visits during the study period (Table 3). Broken down by modality group, those using hybrid care averaged approximately 8 times the number of visits as those using only telehealth and approximately 3 times the number of visits as those receiving only in-person care (19, 2, and 6 visits, respectively).

Table . Number of visits by treatment modality and behavioral health diagnosis.

Diagnosis	Visits, mean (SD)	Breakdown by treatment modality, mean (SD)			ANOVA	
		Both in-person and telehealth	In person only	Telehealth only	*P* value	η^2
All diagnoses	10.81 (13.79)	19.74 (16.57)	5.94 (8.58)	2.43 (4.03)	<.01	0.24
Acute stress disorder	2.09 (3.14)	2.55 (3.82)	1.78 (2.49)	1.38 (2.35)	<.01	0.02
Adjustment disorder	6.67 (8.50)	9.92 (10.49)	4.43 (5.71)	2.53 (3.92)	<.01	0.11
Alcohol use disorder	14.26 (15.54)	18.51 (16.71)	10.10 (12.92)	1.93 (2.88)	<.01	0.08
Attention-deficit/hyperactivity disorder	8.27 (8.76)	10.93 (9.63)	4.18 (4.86)	2.27 (2.62)	<.01	0.15
Generalized anxiety disorder	6.67 (9.04)	8.27 (10.29)	4.43 (6.28)	4.58 (6.58)	<.01	0.04
Insomnia	3.43 (4.31)	4.83 (5.46)	2.49 (2.89)	1.42 (1.22)	<.01	0.08
Major depressive disorder	8.32 (11.01)	10.41 (12.38)	5.17 (7.50)	4.28 (5.78)	<.01	0.06
Panic disorder	4.42 (6.57)	5.53 (7.68)	3.04 (4.40)	3.50 (6.40)	<.01	0.04
Posttraumatic stress disorder	12.02 (13.94)	15.35 (15.40)	7.23 (9.60)	4.15 (6.44)	<.01	0.09
Substance use disorder	8.95 (11.01)	11.50 (12.36)	5.98 (8.16)	1.77 (1.89)	<.01	0.07
Suicidal behavior	1.76 (2.11)	1.91 (2.37)	1.56 (1.68)	1.00 (0[a])	<.01	0.01
Suicidal ideation	2.47 (3.28)	2.85 (3.89)	2.18 (2.66)	1.55 (1.02)	<.01	0.01

[a]SD=0, as all 8 patients attended 1 visit related to suicidal behavior.

The conditions associated with the highest average number of visits included AUD (average of 14 visits; *SD* =15.54), PTSD (average of 12 visits; *SD* =13.94), as well as ADHD, MDD, and SUD (average of 8‑9 visits; *SDs* =8.76-11/01). Insomnia, suicidal behavior, suicidal ideation, and ASD showed the fewest number of visits (2‑3 on average; *SDs* =2.11-4.31). Although all ANOVA tests indicated statistically significant differences in the distribution of care modality within each condition, only 7 of the 12 conditions demonstrated a medium or large effect size (as defined by an η^2 statistic between 0.06 and 0.13 and ≥0.14, respectively). Those seeking care for adjustment disorder, AUD, ADHD, insomnia, MDD, PTSD, and SUD were more likely to receive hybrid care than either in-person or telehealth care alone, as indicated by η^2 statistics and subsequent Tukey tests.

Time Interval Between Visits

Overall, the average interval between behavioral health visits was 67 days (*SD* =138.81; Table 4). In-person–only visits had the longest intervals (70 d; *SD* =153.95), followed by hybrid visits (64 d; *SD* =120.00) and telehealth visits (59 d; *SD* =139.18). Across specific behavioral health conditions, average intervals were longest among those with ASD, ADHD, and insomnia (79‑89 d; *SDs* =116.65-166.60), and the shortest among those with AUD, SUD, and suicidal behavior (25‑38 d; *SDs* =60.93-99.85). In all but 3 conditions (AUD, ADHD, and SUD), those who received only in-person care had the longest average intervals between visits. In 6 of the 12 conditions (ASD, AUD, ADHD, insomnia, PTSD, and SUD), those who received hybrid care of both in-person and telehealth had the shortest interval between visits. However, no models produced a medium or large effect size. Suicidal behavior did not present a significant difference between modalities.

Table . Number of days between visits by treatment modality and behavioral health diagnosis.

Diagnosis	Days between visits, mean (SD)	Breakdown by treatment modality, mean (SD)			ANOVA	
		Both in-person and telehealth	In person only	Telehealth only	*P* value	η^2
All diagnoses	67.07 (138.81)	63.57 (120.00)	70.54 (153.95)	59.12 (139.18)	<.01	0.00
Acute stress disorder	79.05 (157.14)	68.09 (122.39)	91.71 (188.61)	74.80 (151.53)	<.01	0.01
Adjustment disorder	62.33 (127.20)	59.19 (107.54)	65.73 (144.06)	53.38 (131.39)	<.01	0.00
Alcohol use disorder	38.77 (99.85)	33.28 (76.34)	45.89 (123.29)	59.84 (169.24)	<.01	0.00
Attention-deficit/hyperactivity disorder	80.48 (116.65)	75.92 (97.60)	89.46 (146.93)	98.13 (163.71)	<.01	0.00
Generalized anxiety disorder	65.39 (118.25)	58.24 (98.66)	79.27 (146.02)	38.19 (92.95)	<.01	0.01
Insomnia	89.53 (166.60)	75.66 (137.03)	105.31 (193.44)	100.51 (194.46)	<.01	0.01
Major depressive disorder	50.99 (94.97)	45.45 (77.49)	61.68 (120.21)	35.35 (79.00)	<.01	0.01
Panic disorder	64.65 (125.36)	55.19 (95.16)	80.45 (161.05)	45.42 (128.93)	<.01	0.01
Posttraumatic stress disorder	52.92 (104.80)	47.05 (83.30)	63.53 (133.41)	48.34 (133.14)	<.01	0.01
Substance use disorder	30.12 (72.07)	26.66 (58.88)	35.34 (88.21)	42.69 (96.45)	<.01	0.00
Suicidal behavior	25.16 (60.93)	22.96 (32.48)	29.02 (91.40)	22.25 (5.30)	.09	0.00
Suicidal ideation	59.61 (108.08)	50.36 (76.22)	70.47 (134.66)	25.01 (40.29)	<.01	0.01

Discussion

Principal Findings

This study explored the modality of behavioral health care—telehealth, in-person, and hybrid care—delivered to active duty service members within the MHS from 2015 to 2021. This 6-year period spanned from when telehealth was seldom used before the COVID-19 pandemic [7,51,52], during the pandemic when restrictions led to a surge in telehealth use, and after the most stringent pandemic-related restrictions were lifted. During the study period, most service members (60%) received only in-person care, a sizable minority (36%) received a hybrid of telehealth and in-person care, and few (4%) received only telehealth. The higher proportion of in-person and hybrid care may be influenced by the observation period, which consisted of mostly prepandemic years when telehealth was seldom used within the MHS [7,51,52]. The modality of behavioral health care was also examined by demographic characteristics and behavioral health diagnoses. Significant demographic differences emerged showing that women were more likely to use a hybrid of in-person and telehealth care, whereas men more frequently used in-person services alone. Soldiers and airmen used hybrid care more often compared with Marines, while those in the Coast Guard more commonly used telehealth alone. Finally, officers, warrant officers, and senior enlisted members were more likely than junior-enlisted members to use telehealth, both on its own and in combination with in-person care.

Regarding behavioral health diagnoses, service members with ASD, adjustment disorder, insomnia, and suicidal ideation had the highest use of only in-person care and the lowest use of hybrid care. Service members with ADHD, GAD, MDD, PTSD, and suicidal behavior were most likely to use hybrid care and least likely to use only in-person care. Those with AUD, SUD, ADHD, PTSD, MDD, suicidal ideation, and suicidal behavior were least likely to use only telehealth. For 8 of the 12 behavioral health conditions of interest (AUD, ADHD, GAD, MDD, panic disorder, PTSD, SUD, and suicidal behavior), service members were more likely to receive hybrid care, whereas the other 4 conditions (ASD, adjustment disorder, insomnia, and suicidal ideation) were more likely to be associated with in-person care. Although these demographic and diagnostic findings were statistically significant and showed patterns of behavioral health care delivery, effect sizes were small.

Study analyses also compared the delivery modalities in terms of number of visits and between-visit intervals as care metrics. Service members using hybrid care averaged approximately 8 times the number of visits as those using only telehealth and 3 times the number of visits as those receiving in-person care. Specifically, service members who received hybrid care averaged 19 visits (*SD* =16.57), those who received in-person care only averaged 6 visits (*SD* =8.58), and those who received telehealth care only averaged 2 visits (*SD* =4.03). While a sufficient dose of psychotherapy can range depending on clinical factors, such as symptom severity or comorbidities, even a minimally sufficient dose of 9 sessions (such as for PTSD) [55], suggests that only the hybrid group met this threshold.

The average between-visit interval across behavioral health diagnoses was 67 days. This interval exceeds the recommended and commonly evaluated frequencies of once or twice weekly sessions for cognitive behavioral therapies [56,57]. Longer time between sessions was associated with increased dropout among service members in treatment for PTSD [58]. However, it should be noted that not all behavioral health visits were for psychotherapy, and some conditions (eg, ADHD) may be successfully treated with fewer sessions of medication management. Further exploring the frequency of behavioral health care use within the MHS is critical, as it could significantly affect service members' behavioral health and operational readiness through relevant behavioral health policy.

For between-visit intervals across delivery modalities, this difference was statistically significant and amounted to approximately 11 days (70 for in-person only vs 59 for telehealth only). However, it is difficult to determine the extent to which this duration is clinically meaningful. For between-visit intervals across diagnoses, except for 3 (AUD, ADHD, and SUD), service members who received only in-person care had the longest average intervals between visits. Conversely, in 6 of the 12 conditions (ASD, AUD, ADHD, insomnia, PTSD, and SUD), those who received a hybrid of both in-person and telehealth care had the shortest interval between visits. Among specific behavioral health conditions, the longest average intervals were among service members with ASD, ADHD, and insomnia. The shortest intervals were observed among service members with suicidal behavior, SUD, and AUD, which aligns with clinical necessity, as these are presenting concerns often requiring urgent care due to safety risks. One condition, suicidal behavior, did not have a significant difference in the length between visits among the treatment modalities. This may be due, in part, to the small sample size of those with suicidal behavior; for example, there were 8 service members who received only telehealth care.

Comparison With Prior Work

Prior research revealed increased telehealth care use in both the MHS and civilian hospital settings immediately following the onset of the COVID-19 pandemic [4-7,32]. However, the use of a hybrid of telehealth and in-person care has seldom been explored. In a study that examined modalities of treatment delivery, most service members with PTSD, depression, or SUD who initiated behavioral health care early in the pandemic received a hybrid of telehealth and in-person visits [8].

This study adds to the existing literature in several ways. First, the use of telehealth, in-person, and hybrid care was explored over a 6-year period ending in September 2021. In contrast, Hepner et al [8] used corresponding 6-month observation periods (April to September) in 2019 and 2020. The selected time points between these 2 studies highlight different aspects of the data. For example, during this 6-year observation period that included years before the COVID-19 pandemic, in-person behavioral health care was the most common mode of treatment delivery (60%), whereas in the early months following the onset of the pandemic, a hybrid mode of delivery was most frequently received (50% - 56%) [8]. Second, this study uniquely explored whether treatment delivery modality differed across 12 behavioral health conditions of interest. This research question is distinct from that addressed by Hepner et al [8], which determined visits by PTSD, depression, and SUD diagnoses between pre- and post-pandemic periods.

Finally, this work explored care metrics through the number of visits received and intervals between visits across behavioral health conditions and by delivery modality, which showed both similarities and differences with existing research. This study showed similar findings to those of Cozzens [31] regarding reduced time to access care for telehealth compared with in-person visits. The care metrics in this study varied from those explored by Hepner et al [8], which focused on treatment initiation and transitions of care by the 3 diagnoses of interest during the pre- and post-pandemic periods rather than by both diagnosis and delivery modality. In sum, this study complements and builds on the existing literature by extending the postpandemic period and determining the use of delivery modality across a wide array of behavioral health diagnoses, which can inform ongoing health care delivery within the MHS.

Limitations

There are several limitations that should be considered when interpreting the results of this research. Study data were not based on gold standard, diagnostic assessments but rather, were derived from diagnostic and procedural codes documented in electronic medical records, which may be subject to factors such as coding errors, provider knowledge, and the extent of symptoms discussed in an appointment. Specific to telehealth, there was evolving guidance regarding how providers should code for telehealth services in the MHS that could have contributed to variability [8]. Although guidance issued directly to behavioral health providers during the pandemic period was obtained by the authors and reviewed for the extraction of relevant codes, the validity of these codes over time cannot be ascertained. Health service and policy researchers have proposed guidance for providers regarding the coding of telehealth services in the MHS to improve data accuracy [8,31]. Additionally, data were only available for service members who received behavioral health care that was reimbursed by TRICARE, and findings may not extend beyond this population. Separate courses of treatment could not be determined from medical record data and, with the 6-year period, may result in longer average between-visit intervals. Data from medical records indicate health care use and may not represent the care preferences of service members or satisfaction with care received. Finally, this study captured trends over an observation period that included a critical period in telehealth use within the MHS; however, it does not reflect current patterns of modality use or those since the declassification of the COVID-19 pandemic as a public health emergency [59], thus necessitating ongoing research efforts.

Conclusions

Behavioral health conditions can adversely affect service members and operational readiness. Offering options beyond in-person behavioral health care may improve access to care, as study results demonstrated. Collectively, findings from the

6-year observation period showed that telehealth was commonly used in combination with in-person care. Furthermore, telehealth was related to more behavioral health care visits and the least amount of time between visits, highlighting the advantages of offering telehealth as an option within the MHS health care landscape. Although study findings support telehealth as an option for treatment delivery, it may not be ideally suited for all service members or in all situations [32], and in-person or hybrid care delivery may be preferred by a patient or deemed more clinically appropriate by a provider. Given options for care delivery within the MHS, it is recommended that treatment modality be selected based on patient preference and shared decision-making [49,50]. Additionally, providing ongoing flexibility, regularly reassessing preferences, and personalizing treatment are important aspects for the delivery of optimal behavioral health care to service members [22,50], along with the infrastructure and policies to support these practices [42]. This study contributes novel information about behavioral health treatment delivery within the MHS, but further research is needed to explore service member preferences for delivery modality (telehealth, in-person, and hybrid care) and how preferences align with care received and treatment outcomes.

Acknowledgments

The authors would like to acknowledge the CHAMPS database management of Thierry Nedellec, editorial support of Michelle Stoia, and reference and formatting assistance of Hayley Myers.

Funding

KHW, LHG, JAL, JFB, and SYC acknowledge funding support of the Military Health System Research Program. EAS and HJJ acknowledge the funding support of S21MD010690 (SDSU HealthLINK Endowment).

Disclaimer

KHW is an employee of the US Government. This work was prepared as part of her official duties. Title 17, U.S.C. §105 provides that copyright protection under this title is not available for any work of the US Government. Title 17, U.S.C. §101 defines a US Government work as work prepared by a military service member or employee of the US Government as part of that person's official duties. Report No. 25 - 62 was supported by the Military Health System Research Program under Project ID 62387. The views expressed in this work are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the US Government.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to personally identifiable information regulations, but they may be made available by the corresponding author upon reasonable request and approval by the Naval Health Research Center Institutional Review Board or Privacy Office.

Authors' Contributions

Conceptualization: KHW, LHG, EAS Data curation: JFB Formal analysis: JAL, HJJ, SYC Funding acquisition: KHW, LHG, EAS Investigation: KHW, LHG, EAS Methodology: KHW, EAS Project administration: KHW Resources: KHW, EAS Supervision: KHW, LHG, EAS Validation: JAL, JFB, SYC Writing – original draft: KHW, JAL, SYC Writing – review and editing: LHG, HJJ, JFB, EAS

Conflicts of Interest

None declared.

Multimedia Appendix 1
Medical billing codes denoting behavioral health treatment, telehealth treatment, and behavioral health diagnosis.
[DOCX File, 17 KB - mental_v13i1e83809_app1.docx]

References

1. ATA's standardized telehealth terminology and policy language for states on medical practice. American Telemedicine Association. 2020 Sep 21. URL: https://www.americantelemed.org/wp-content/uploads/2020/10/ATA-_Medical-Practice-10-5-20.pdf [accessed 2024-11-15]
2. Breslau J, Finucane ML, Locker AR, Baird MD, Roth EA, Collins RL. A longitudinal study of psychological distress in the United States before and during the COVID-19 pandemic. Prev Med 2021 Feb;143:106362. [doi: 10.1016/j.ypmed.2020.106362] [Medline: 33388325]
3. Lee JEC, Bennett C, Bennett N, et al. Assessing military mental health during the pandemic: a five country collaboration. Curr Psychiatry Rep 2025 Dec;27(12):733-742. [doi: 10.1007/s11920-024-01522-3] [Medline: 39394493]

4. Patel SY, Mehrotra A, Huskamp HA, Uscher-Pines L, Ganguli I, Barnett ML. Trends in outpatient care delivery and telemedicine during the COVID-19 pandemic in the US. JAMA Intern Med 2021 Mar 1;181(3):388-391. [doi: 10.1001/jamainternmed.2020.5928] [Medline: 33196765]
5. Gilder T, Banaag A, Madsen C, Koehlmoos TP. Trends in telehealth care during the COVID-19 pandemic for the Military Health System. Telemed Rep 2023;4(1):147-155. [doi: 10.1089/tmr.2022.0042] [Medline: 37771698]
6. Clark L, Fan M, Stahlman S. Surveillance of mental and behavioral health care utilization and use of telehealth, active component, U.S. Armed Forces, 1 January 2019-30 September 2020. MSMR 2021 Aug 1;28(8):22-27. [Medline: 34622900]
7. Glassman LH, Schmied EA, Jun HJ, Bonkowski JF, Levine JA, Walter KH. Telebehavioral health care utilization among U.S. military personnel before and during the COVID-19 pandemic. Mil Med 2025 Sep 1;190(Supplement_2):678-685. [doi: 10.1093/milmed/usaf309] [Medline: 40984144]
8. Hepner KA, Roth CP, Sousa JL, Ruder T, Brown RA, Parast L, et al. Behavioral health care delivery following the onset of the COVID-19 pandemic: utilization, telehealth, and quality of care for service members with PTSD, depression, or substance use disorder. RAND Corporation. 2023. URL: https://www.rand.org/pubs/research_reports/RRA421-3.html [accessed 2024-11-15]
9. Heyworth L, Shah N, Galpin K. 20 years of telehealth in the Veterans Health Administration: taking stock of our past and charting our future. J Gen Intern Med 2024 Feb;39(Suppl 1):5-8. [doi: 10.1007/s11606-024-08617-w] [Medline: 38378981]
10. Fischer SH, Ray KN, Mehrotra A, Bloom EL, Uscher-Pines L. Prevalence and characteristics of telehealth utilization in the United States. JAMA Netw Open 2020 Oct 1;3(10):e2022302. [doi: 10.1001/jamanetworkopen.2020.22302] [Medline: 33104208]
11. Ashwick R, Turgoose D, Murphy D. Exploring the acceptability of delivering cognitive processing therapy (CPT) to UK veterans with PTSD over Skype: a qualitative study. Eur J Psychotraumatol 2019;10(1):1573128. [doi: 10.1080/20008198.2019.1573128] [Medline: 30774784]
12. Acierno R, Knapp R, Tuerk P, et al. A non-inferiority trial of Prolonged Exposure for posttraumatic stress disorder: In person versus home-based telehealth. Behav Res Ther 2017 Feb;89:57-65. [doi: 10.1016/j.brat.2016.11.009] [Medline: 27894058]
13. Cully JA, Jameson JP, Phillips LL, Kunik ME, Fortney JC. Use of psychotherapy by rural and urban veterans. J Rural Health 2010;26(3):225-233. [doi: 10.1111/j.1748-0361.2010.00294.x] [Medline: 20633090]
14. Seal KH, Abadjian L, McCamish N, Shi Y, Tarasovsky G, Weingardt K. A randomized controlled trial of telephone motivational interviewing to enhance mental health treatment engagement in Iraq and Afghanistan veterans. Gen Hosp Psychiatry 2012;34(5):450-459. [doi: 10.1016/j.genhosppsych.2012.04.007] [Medline: 22632925]
15. Wierwille JL, Pukay-Martin ND, Chard KM, Klump MC. Effectiveness of PTSD telehealth treatment in a VA clinical sample. Psychol Serv 2016 Nov;13(4):373-379. [doi: 10.1037/ser0000106] [Medline: 27657798]
16. Lustig TA. The Role of Telehealth in an Evolving Health Care Environment: Workshop Summary: The National Academies Press; 2012. [doi: 10.17226/13466]
17. Hall A, Leech J, Schmeider L, Swayze M, Saenz J, Currier G. Mental health evacuation rate in USCENTCOM. Mil Med 2024 Aug 19;189(Suppl 3):18-20. [doi: 10.1093/milmed/usae032] [Medline: 39160801]
18. Hall A, Olsen C, Gomes J, Bajjani-Gebara J, Meyers E, Wilson R. Relative Risk of All-Cause Medical Evacuation for Behavioral Health Conditions in U.S. Central Command. Mil Med 2024 Jan 23;189(1-2):e279-e284. [doi: 10.1093/milmed/usad306] [Medline: 37552646]
19. Chen PV, Helm A, Fletcher T, et al. Seeing the Value of Video: A Qualitative Study on Patient Preference for Using Video in a Veteran Affairs Telemental Health Program Evaluation. Telemed Rep 2021;2(1):156-162. [doi: 10.1089/tmr.2021.0005] [Medline: 35720740]
20. Acierno R, Gros DF, Ruggiero KJ, et al. Behavioral activation and therapeutic exposure for posttraumatic stress disorder: a noninferiority trial of treatment delivered in person versus home-based telehealth. Depress Anxiety 2016 May;33(5):415-423. [doi: 10.1002/da.22476] [Medline: 26864655]
21. Hoge CW, Castro CA, Messer SC, McGurk D, Cotting DI, Koffman RL. Combat duty in Iraq and Afghanistan, mental health problems, and barriers to care. N Engl J Med 2004 Jul 1;351(1):13-22. [doi: 10.1056/NEJMoa040603] [Medline: 15229303]
22. Jones C, Miguel-Cruz A, Smith-MacDonald L, et al. Virtual trauma–focused therapy for military members, veterans, and public safety personnel with posttraumatic stress injury: systematic scoping review. JMIR Mhealth Uhealth 2020 Sep 21;8(9):e22079. [doi: 10.2196/22079] [Medline: 32955456]
23. Madsen C, Poropatich R, Koehlmoos TP. Telehealth in the Military Health System: impact, obstacles, and opportunities. Mil Med 2023 Mar 6;188(Suppl 1):15-23. [doi: 10.1093/milmed/usac207] [Medline: 36882030]
24. Morland LA, Mackintosh MA, Greene CJ, et al. Cognitive processing therapy for posttraumatic stress disorder delivered to rural veterans via telemental health: a randomized noninferiority clinical trial. J Clin Psychiatry 2014 May;75(5):470-476. [doi: 10.4088/JCP.13m08842] [Medline: 24922484]
25. Stecker T, Shiner B, Watts BV, Jones M, Conner KR. Treatment-seeking barriers for veterans of the Iraq and Afghanistan conflicts who screen positive for PTSD. Psychiatr Serv 2013 Mar 1;64(3):280-283. [doi: 10.1176/appi.ps.001372012] [Medline: 23450385]

26. Bellanti DM, Kelber MS, Workman DE, Beech EH, Belsher BE. Rapid review on the effectiveness of telehealth interventions for the treatment of behavioral health disorders. Mil Med 2022 May 3;187(5-6):e577-e588. [doi: 10.1093/milmed/usab318] [Medline: 34368853]
27. McClellan MJ, Osbaldiston R, Wu R, et al. The effectiveness of telepsychology with veterans: a meta-analysis of services delivered by videoconference and phone. Psychol Serv 2022 May;19(2):294-304. [doi: 10.1037/ser0000522] [Medline: 33539135]
28. Turgoose D, Ashwick R, Murphy D. Systematic review of lessons learned from delivering tele-therapy to veterans with post-traumatic stress disorder. J Telemed Telecare 2018 Oct;24(9):575-585. [doi: 10.1177/1357633X17730443] [Medline: 28958211]
29. Olden M, Wyka K, Cukor J, et al. Pilot study of a telehealth-delivered medication-augmented exposure therapy protocol for PTSD. J Nerv Ment Dis 2017 Feb;205(2):154-160. [doi: 10.1097/NMD.0000000000000563] [Medline: 27441461]
30. Ziemba SJ, Bradley NS, Landry LAP, Roth CH, Porter LS, Cuyler RN. Posttraumatic stress disorder treatment for Operation Enduring Freedom/Operation Iraqi Freedom combat veterans through a civilian community–based telemedicine network. Telemed J E Health 2014 May;20(5):446-450. [doi: 10.1089/tmj.2013.0312] [Medline: 24617961]
31. Cozzens FJ. A cost benefit analysis on the utilization of telemedicine services for mental health. EMBA capstone project report. Naval Postgraduate School. 2021.
32. Lancaster SL, Linkh DJ, Lawless CE, Renno S. Comparison of telehealth and in-person mental health care in military veterans and active-duty service members. Psychol Serv 2025 May;22(2):215-220. [doi: 10.1037/ser0000868] [Medline: 38780556]
33. Donelan K, Barreto EA, Sossong S, et al. Patient and clinician experiences with telehealth for patient follow-up care. Am J Manag Care 2019 Jan;25(1):40-44. [Medline: 30667610]
34. Murphy D, Turgoose D. Evaluating an internet-based video cognitive processing therapy intervention for veterans with PTSD: a pilot study. J Telemed Telecare 2020 Oct;26(9):552-559. [doi: 10.1177/1357633X19850393] [Medline: 31208264]
35. Shore P, Goranson A, Ward MF, Lu MW. Meeting veterans where they're @: a VA Home-Based Telemental Health (HBTMH) pilot program. Int J Psychiatry Med 2014;48(1):5-17. [doi: 10.2190/PM.48.1.b] [Medline: 25354923]
36. Whealin JM, King L, Shore P, Spira JL. Diverse veterans' pre- and post-intervention perceptions of home telemental health for posttraumatic stress disorder delivered via tablet. Int J Psychiatry Med 2017 Jan;52(1):3-20. [doi: 10.1177/0091217417703291] [Medline: 28486881]
37. Yuen EK, Gros DF, Price M, et al. Randomized controlled trial of home-based telehealth versus in-person prolonged exposure for combat-related PTSD in veterans: preliminary results. J Clin Psychol 2015 Jun;71(6):500-512. [doi: 10.1002/jclp.22168] [Medline: 25809565]
38. Kichloo A, Albosta M, Dettloff K, et al. Telemedicine, the current COVID-19 pandemic and the future: a narrative review and perspectives moving forward in the USA. Fam Med Community Health 2020 Aug;8(3):e000530. [doi: 10.1136/fmch-2020-000530] [Medline: 32816942]
39. Mani V, Pomer A, Madsen C, et al. Filling the gaps in the pandemic response: impact of COVID-19 on telehealth in the Military Health System. Telemed J E Health 2024 May;30(5):1443-1449. [doi: 10.1089/tmj.2023.0478] [Medline: 38126844]
40. Appleton R, Williams J, Vera San Juan N, et al. Implementation, adoption, and perceptions of telemental health during the COVID-19 pandemic: systematic review. J Med Internet Res 2021 Dec 9;23(12):e31746. [doi: 10.2196/31746] [Medline: 34709179]
41. Batastini AB, Paprzycki P, Jones ACT, MacLean N. Are videoconferenced mental and behavioral health services just as good as in-person? A meta-analysis of a fast-growing practice. Clin Psychol Rev 2021 Feb;83:101944. [doi: 10.1016/j.cpr.2020.101944] [Medline: 33227560]
42. Boykin DM, Keegan F, Thompson KE, Voelkel E, Lindsay JA, Fletcher TL. Video to home delivery of evidence-based psychotherapy to veterans with posttraumatic stress disorder. Front Psychiatry 2019;10:893. [doi: 10.3389/fpsyt.2019.00893] [Medline: 31920747]
43. Lythreatis S, Singh SK, El-Kassar AN. The digital divide: a review and future research agenda. Technol Forecast Soc Change 2022 Feb;175:121359. [doi: 10.1016/j.techfore.2021.121359]
44. Kelber MS, Smolenski DJ, Boyd C, et al. Evidence-based telehealth interventions for post-traumatic stress disorder, depression, and anxiety: a systematic review and meta-analysis. J Telemed Telecare 2025 Jul;31(6):757-767. [doi: 10.1177/1357633X231224491] [Medline: 38254285]
45. Luxton DD, Pruitt LD, Wagner A, Smolenski DJ, Jenkins-Guarnieri MA, Gahm G. Home-based telebehavioral health for U.S. military personnel and veterans with depression: a randomized controlled trial. J Consult Clin Psychol 2016 Nov;84(11):923-934. [doi: 10.1037/ccp0000135] [Medline: 27599225]
46. Mohr DC, Ho J, Duffecy J, et al. Effect of telephone-administered vs face-to-face cognitive behavioral therapy on adherence to therapy and depression outcomes among primary care patients: a randomized trial. JAMA 2012 Jun 6;307(21):2278-2285. [doi: 10.1001/jama.2012.5588] [Medline: 22706833]
47. Study and report on increasing telehealth services across armed forces. Department of Defense. 2022. URL: https://www.health.mil/Reference-Center/Reports/2022/10/21/Study-and-Report-on-Increasing-Telehealth-Services-across-Armed-Forces [accessed 2024-11-15]

48. Strategic Plan: Fiscal Years 2025-2030. Defense Health Agency. 2025. URL: https://dha.mil/-/media/Project/Documents/DHA_StrategicPlan_20240910_FINAL_2pager.pdf [accessed 2026-11-2]
49. Morland LA, Wells SY, Glassman LH, Greene CJ, Hoffman JE, Rosen CS. Advances in PTSD treatment delivery: review of findings and clinical considerations for the use of telehealth interventions for PTSD. Curr Treat Options Psychiatry 2020;7(3):221-241. [doi: 10.1007/s40501-020-00215-x] [Medline: 32837831]
50. Schlief M, Saunders KRK, Appleton R, et al. Synthesis of the evidence on what works for whom in telemental health: rapid realist review. Interact J Med Res 2022 Sep 29;11(2):e38239. [doi: 10.2196/38239] [Medline: 35767691]
51. Health care: telehealth and remote patient monitoring use in medicare and selected federal programs (GAO-17-365). Government Accountability Office. 2017. URL: https://www.gao.gov/assets/gao-17-365.pdf [accessed 2024-11-15]
52. Hepner KA, Brown RA, Roth CP, Ruder T, Pincus HA. Behavioral health care in the military health system: access and quality for remote service members. RAND Corporation. 2021. URL: https://www.rand.org/pubs/research_reports/RR2788.html [accessed 2024-11-15]
53. Walter KH, Levine JA, Highfill-McRoy RM, Navarro M, Thomsen CJ. Prevalence of posttraumatic stress disorder and psychological comorbidities among US active duty service members, 2006–2013. J Trauma Stress 2018 Dec;31(6):837-844. [doi: 10.1002/jts.22337] [Medline: 30398680]
54. International Classification of Diseases, Tenth Revision. World Health Organization. 1993. URL: https://icd.who.int/browse10/2019/en [accessed 2025-12-29]
55. Harper KL, Lee DJ, Moshier S, Zweig I, Keane TM, Marx BP. Is adequate dose adequate? An examination of the impact of psychotherapy on posttraumatic stress disorder symptoms utilizing Veterans Health Administration medical records. Psychol Serv 2025 Feb;22(1):167-176. [doi: 10.1037/ser0000830] [Medline: 38271024]
56. Ciharova M, Karyotaki E, Miguel C, et al. Amount and frequency of psychotherapy as predictors of treatment outcome for adult depression: a meta-regression analysis. J Affect Disord 2024 Aug 15;359:92-99. [doi: 10.1016/j.jad.2024.05.070] [Medline: 38777269]
57. Gutner CA, Suvak MK, Sloan DM, Resick PA. Does timing matter? Examining the impact of session timing on outcome. J Consult Clin Psychol 2016 Dec;84(12):1108-1115. [doi: 10.1037/ccp0000120] [Medline: 27213491]
58. Fleming CJE, Hawrilenko M, Wachen JS, et al. It's about time: examining the role of session timing in cognitive processing therapy in active duty military personnel. Journal of Behavioral and Cognitive Therapy 2020 Sep;30(3):231-239. [doi: 10.1016/j.jbct.2020.04.001]
59. Fact sheet: end of the COVID-19 public health emergency. US Department of Health and Human Services. 2023. URL: https://www.hhs.gov/about/news/2023/05/09/fact-sheet-end-of-the-covid-19-public-health-emergency.html [accessed 2024-11-15]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
ASD: acute stress disorder
AUD: alcohol use disorder
GAD: generalized anxiety disorder
ICD-10: International Classification of Diseases, 10th Revision
MDD: major depressive disorder
MDR: Military Health System Data Repository
MHS: Military Health System
PTSD: posttraumatic stress disorder
SUD: substance use disorder

Edited by C Huber; submitted 09.Sep.2025; peer-reviewed by NH Corry, TO Walton; revised version received 15.Nov.2025; accepted 08.Dec.2025; published 05.Jan.2026.

Please cite as:
Walter KH, Glassman LH, Levine JA, Jun HJ, Bonkowski JF, Chung SY, Schmied EA
Telebehavioral Health, In-Person, and Hybrid Modalities of Treatment Delivery Among US Service Members: Longitudinal Observational Study
JMIR Ment Health 2026;13:e83809
URL: https://mental.jmir.org/2026/1/e83809
doi:10.2196/83809

Trends in the Implementation of the Cyberchondria Severity Scale: Bibliometric Analysis

Adam C Powell[1,2], PhD; Cayetana Calderon-Smith[3], MSEd, MPhilEd

[1]Payer+Provider Syndicate, 20 Oakland Ave, Newton, MA, United States

[2]College of Population Health, Thomas Jefferson University, Philadelphia, PA, United States

[3]Therapy 2.0, Bala Cynwyd, PA, United States

Corresponding Author:
Adam C Powell, PhD
Payer+Provider Syndicate, 20 Oakland Ave, Newton, MA, United States

Abstract

Background: Cyberchondria, a combination of the words "cyber" and "hypochondriasis," is a condition that is receiving increasing attention from clinicians and researchers globally. Researchers are currently using multiple instruments to quantify it. Furthermore, the instruments have been translated into multiple languages.

Objective: This study aimed to examine the extent to which researchers are measuring cyberchondria using the 33-item Cyberchondria Severity Scale (CSS) and its 12-item abbreviated version, the CSS-12. It also examined the relative use of cyberchondria instruments in different languages.

Methods: PubMed and PsycInfo were searched for articles published between May 1, 2019, and December 31, 2024, featuring the term "cyberchondria" in the title. Included articles mentioned the CSS, were empirical studies, and were in English. Each article was categorized by the CSS version, publication year, and language of instrument implementation. Fisher exact tests were used to assess associations, and the Spearman rank correlation coefficient was used to evaluate trend monotonicity.

Results: Among the 117 articles included in the analysis, 42 (35.9%) used the CSS, 38 (32.5%) used the CSS-12, and the remaining 37 (31.6%) used unknown or modified versions. Although CSS-12 use began with its introduction in 2019, there was no significant association between publication year and instrument choice (P=.84). Unadjusted analysis found that the relationship between year and the percentage of articles using the CSS-12 showed a statistically significant monotonic trend (ρ=0.89; P=.02). This finding was not significant after applying a Bonferroni correction. However, there was a significant association between the language of the instrument and the CSS version used (P<.001).

Conclusions: From 2019 to 2024, both the CSS and CSS-12 continued to be used. The CSS-12 offers benefits such as brevity and the removal of reverse-keyed items, while the original CSS remains useful for studies that require the mistrust of medical professionals subscale. The significant association between language and instrument choice suggests that cultural and linguistic factors impact selection, and instrument choice should be guided by the study's objectives and the constructs of interest.

(JMIR Ment Health 2026;13:e75003) doi:10.2196/75003

KEYWORDS

health anxiety; health information seeking behavior; hypochondriasis; hypochondria; Cyberchondria Severity Scale; CSS; 12-item Cyberchondria Severity Scale; CSS-12

Introduction

Definition and Current Measurement Tools

Cyberchondria is a portmanteau of the words "cyber" and "hypochondriasis." Its measurement was first formalized through the development of the Cyberchondria Severity Scale (CSS) in 2014 [1]. In its original format, the CSS consists of a 33-item questionnaire, grouped into 5 subscales, some of which identify behaviors (ie, compulsiveness, excessiveness, and reassurance seeking) or mental states (ie, distress). An additional subscale, mistrust of medical professionals, has the potential to be problematic, as it may measure a construct that is *different from, but related to*, the other 4 cyberchondria subscales [2].

In response to both the length of the original, long-form CSS and the potential issues surrounding the mistrust of medical professionals subscale, an abbreviated version called the CSS-12 was developed in 2019 [3]. The CSS-12 consists of a 12-item questionnaire containing questions drawn from the original version; however, it does not include any items related to the mistrust of medical professionals. The creators of the original CSS were involved in the development and validation of CSS-12 and thus have implicitly endorsed it.

Since their creation, the CSS and CSS-12 have been used in numerous studies and have become de facto standards for the measurement of cyberchondria. A potential overreliance on the CSS is acknowledged in the literature [4]. Furthermore, the instruments have been translated into other languages and have been extensively used in adapted forms. On this note, in 2016, a German team created the 15-question German version of the instrument, dubbed the CSS-15 [5]. Additional novel instruments have been developed, some of which include the aforementioned mistrust of medical professionals construct [6,7].

Study Aims

There is widespread use of the CSS and CSS-12 and a lack of research comparing their relative use. To address this lacuna, this study aims to provide future researchers with greater understanding of the extent to which each version is used and the degree to which each version is being used in languages other than English. It additionally aims to contribute to the discussion of the various contexts in which inclusion of the mistrust of medical professionals subscale is helpful. To achieve these aims, we conducted a review of the literature to determine the relative frequency with which the CSS and CSS-12 were used and the languages in which they were used. While conducting this review, situations in which noncanonical forms of the questionnaire were used were noted.

Methods

Ethical Considerations

Ethics approval and informed consent were not applicable because this study examined the published literature, rather than human participants.

Search Strategy and Sample Selection

In September 2025, PubMed and PsycInfo were searched for all potentially relevant articles published between May 1, 2019, the date of publication of the article defining the CSS-12, and December 31, 2024, the last day of the most recent calendar year. PubMed is a free tool that searches the archive of biomedical and life sciences journal literature maintained by the United States National Library of Medicine. It may be most accessible to clinical practitioners without institutional access to paywalled sources. PsycInfo is a database of articles administered by the American Psychological Association. Articles likely to be about cyberchondria were initially identified by searching for peer-reviewed, published articles with "cyberchondria" in the title. A pool of articles to evaluate was created by removing the duplicates found by both sources. Articles were excluded if they were replies, corrigenda, letters to the editor, letters from the editor, or not actually published during the search period. Further exclusions were made for articles that were not in English, were reviews, contained conceptual analysis, or did not measure cyberchondria.

Measurement

Each article was reviewed to determine whether it used the original 33-question CSS, the CSS-12, another form of the CSS (eg, the CSS-15 or an author-derived version), or an unknown version. Culturally equivalent translations of the CSS or CSS-12 from English into another language were classified as being the instrument that was translated. The process used to determine the version of the scale used is described in Multimedia Appendix 1. Two variables were created to capture the language of the instrument: 1 variable that categorized studies as having used an instrument with an "unspecified" language if it was not explicitly stated and 1 variable that attempted to infer the language of the instrument used in studies based upon the context in which they were conducted. The process used to ascertain the language used in an article is described in Multimedia Appendix 2.

Articles were additionally classified by year of publication and by the language in which the instrument was implemented. While only English-language articles were considered, articles were written by authorship teams from various nations and, in many cases, reported on empirical research that was not conducted in English. Studies conducted in English-speaking countries were assumed to have used an English version of the instrument, unless explicitly stated otherwise. This assumption was made, as the original implementations of the CSS and CSS-12 were in English.

Analysis

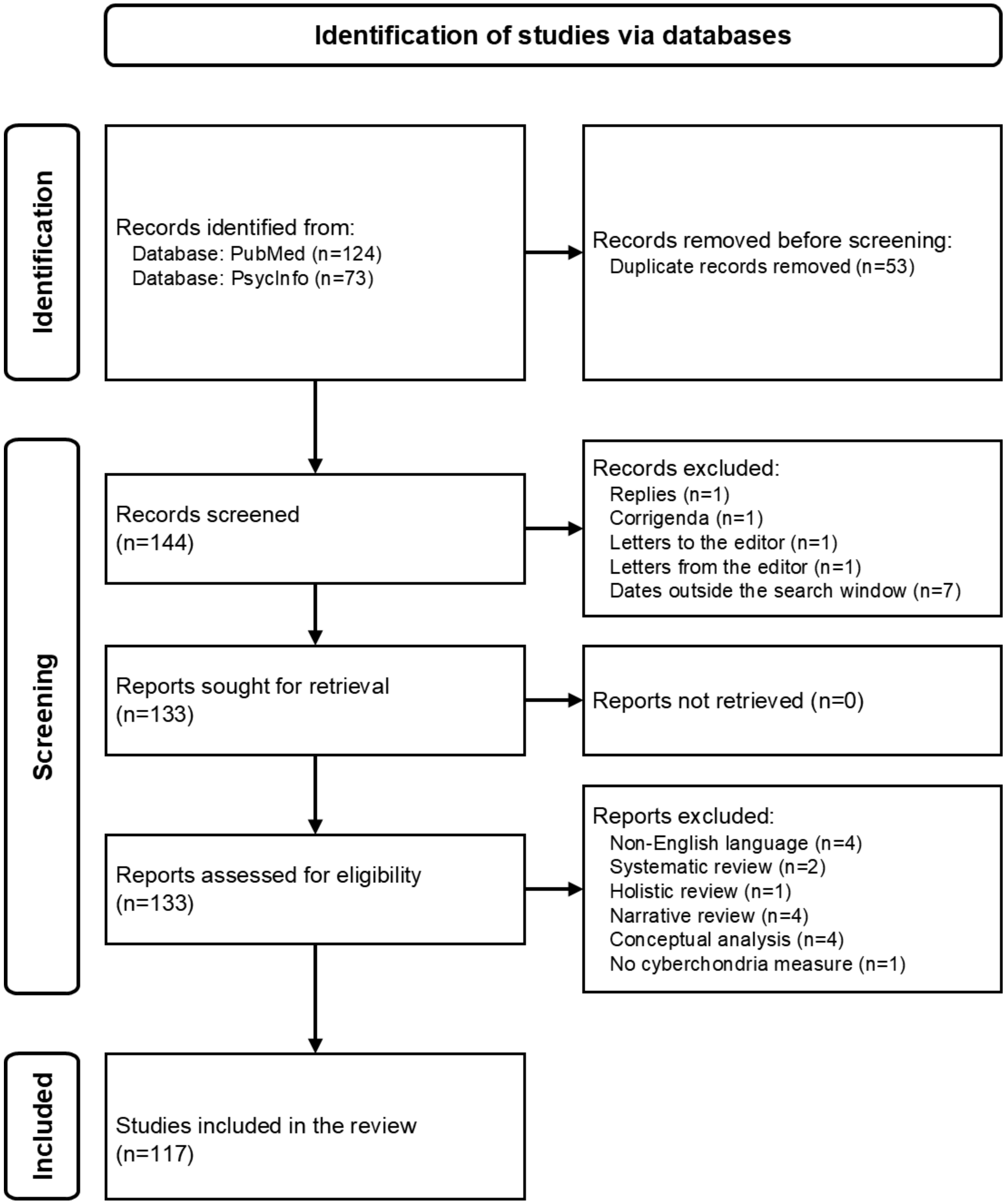
For each year, 2019 to 2024, the number of articles using the CSS, the CSS-12, and other variations of the CSS was determined by reviewing the contents of the articles meeting the sample selection criteria, and results were recorded in a table. If the version of the CSS used could not be determined, it was classified as "unknown." Fisher exact tests were used to assess the relationship between year and the type of CSS instrument used, considering both the totality of the articles and a subset using only the CSS or CSS-12. Spearman rank correlation coefficient was calculated to determine whether there was a trend in the percentage of cyberchondria articles using the CSS-12 that was monotonic. The percentage of cyberchondria articles using the CSS-12 was then plotted by year using a scatterplot.

The sample was examined to determine the language used to assess cyberchondria in each study considered. For each language found in the sample, the number of studies using the CSS, the CSS-12, and author-derived variations of the CSS was determined. Fisher exact tests were run to assess whether a significant association existed between language and CSS implementation used, again considering both the totality of the articles and the subset using only the CSS or CSS-12.

Results

Searching PubMed yielded 124 articles, and searching PsycInfo yielded 73 articles. Of these 197 articles, 144 (73.1%) were unique. The exclusion criteria were applied as shown in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram (Figure 1), leading to 117 (59.4%) studies being included in the review. Of these 117 articles, 42 (35.9%) used the CSS, 38 (32.5%) used the CSS-12, 36 (30.8%) used other instruments, and 1 (0.9%) used an unknown instrument.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram depicting the sample creation process.



As shown in Table 1, while the CSS-12 was introduced in 2019, it took some time for it to gain widespread use after its initial introduction [3]. Only 11% (1/9) of the articles used it in 2020, and only 36% (5/14) of the articles used it in 2021. A Fisher exact test did not identify a significant association between the year of publication and the instrument used (*P*=.84). When a Fisher exact test was run considering only studies that used the CSS or CSS-12 (excluding studies using instruments classified as other and unknown), there was still no significant relationship between the year of publication and the instrument used (*P*=.54). Spearman rank correlation coefficient showed a statistically significant monotonic relationship between the year of publication and the proportion of studies using the CSS-12 (ρ=0.89; *P*=.02). The year in which the greatest proportion of the studies used the CSS-12 was 2024, when 39% (9/23) of the studies used the instrument.

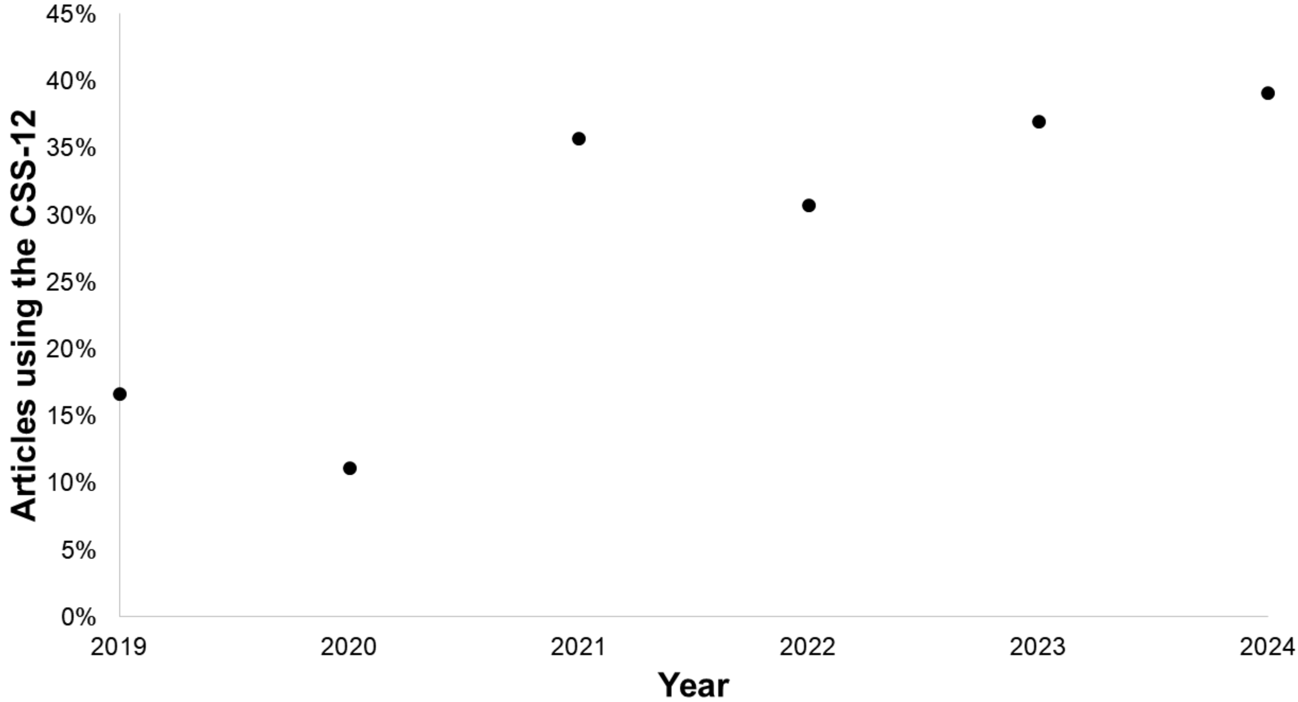
Table . Instrument use by year, 2019 to 2024.

	CSS[a], n (%)	CSS-12[b], n (%)	Other, n (%)	Unknown, n (%)
2019 (n=6)	2 (33.3) [8,9]	1 (16.7) [3]	3 (50) [7,10,11]	0 (0)
2020 (n=9)	5 (55.6) [12-16]	1 (11.1) [17]	3 (33.3) [18-20]	0 (0)
2021 (n=14)	6 (42.9) [21-26]	5 (35.7) [27-31]	3 (21.4) [32-34]	0 (0)
2022 (n=39)	15 (38.5) [35-49]	12 (30.8) [50-61]	11 (28.2) [62-72]	1 (2.6) [73]
2023 (n=26)	6 (23.1) [74-79]	10 (38.5) [80-89]	10 (38.5) [90-99]	0 (0)
2024 (n=23)	8 (34.8) [100-107]	9 (39.1) [108-116]	6 (26.1) [6,117-121]	0 (0)
Grand total (n=117)	42 (35.9)	38 (32.5)	36 (30.8)	1 (0.9)

[a]CSS: Cyberchondria Severity Scale.
[b]CSS-12: 12-item Cyberchondria Severity Scale.

As shown in the scatterplot presented in Figure 2, the only cyberchondria article published in 2019 mentioning the CSS-12 was the article that defined the instrument [3]. Use of the CSS-12 exceeded use of the CSS in 2023 and 2024, but studies using the CSS-12 did not account for the majority of studies due to the various other versions of the instrument that were used.

Figure 2. Percentage of cyberchondria articles using the 12-item Cyberchondria Severity Scale (CSS-12) by year.



As shown in Table 2, among the articles that explicitly stated the language that was used, the CSS saw the greatest adoption in articles that implemented it in Turkish (14/42, 33%), and the CSS-12 saw the greatest adoption in articles that implemented it in Chinese (6/38, 16%) or Turkish (6/38, 16%). The languages for which there were more articles written using the CSS-12 than the CSS were Arabic (4/5, 80%), Chinese (6/9, 67%), Persian (3/4, 75%), Russian (1/1, 100%), Spanish (2/2, 100%), and Serbian (1/2, 50%). A Fisher exact test identified a significant association between the language in which an article implemented its cyberchondria measurement and the instrument used (*P*<.001). When articles using an instrument other than the CSS or CSS-12 were excluded from the analysis, a Fisher exact test likewise identified a significant association between the language in which an article implemented its cyberchondria measurement and the instrument used (*P*=.03).

Table . Instrument use by language (ambiguous cases classified as "unspecified").

Language	CSS[a], n (%)	CSS-12[b], n (%)	Other, n (%)	Unknown, n (%)
Arabic (n=5)	0 (0)	4 (80) [27,80-82]	1 (20) [117]	0 (0)
Chinese (n=9)	2 (22.2) [35,74]	6 (66.7) [28,50,83,108-110]	1 (11.1) [90]	0 (0)
Croatian (n=6)	1 (16.7) [100]	0 (0)	5 (83.3) [7,18,91,92,118]	0 (0)
English (n=18)	7 (38.9) [8,9,12,13,21,22,36]	5 (27.8) [3,84,111-113]	5 (27.8) [10,32,62,93,119]	1 (5.6) [73]
French (n=1)	0 (0)	1 (100) [85]	0 (0)	0 (0)
German (n=3)	1 (33.3) [101]	0 (0)	2 (66.7) [19,63]	0 (0)
Indonesian (n=1)	0 (0)	0 (0)	1 (100) [64]	0 (0)
Italian (n=6)	3 (50) [14,37,38]	3 (50) [29,51,52]	0 (0)	0 (0)
Korean (n=1)	0 (0)	0 (0)	1 (100) [94]	0 (0)
Persian (n=4)	1 (25) [75]	3 (75) [17,30,53]	0 (0)	0 (0)
Polish (n=8)	7 (87.5) [15,23,24,39,102-104]	1 (12.5) [54]	0 (0)	0 (0)
Portuguese (n=1)	1 (100) [40]	0 (0)	0 (0)	0 (0)
Russian (n=1)	0 (0)	1 (100) [55]	0 (0)	0 (0)
Serbian (n=2)	0 (0.0)	1 (50.0) [56]	1 (50.0) [95]	0 (0.0)
Spanish (n=2)	0 (0)	2 (100) [57,86]	0 (0)	0 (0)
Turkish (n=24)	14 (58.3) [25,41-46,76-79,105-107]	6 (25) [31,58-60,87,114]	4 (16.7) [33,65,66,96]	0 (0)
Urdu (n=1)	0 (0)	0 (0)	1 (100) [99]	0 (0)
Unspecified (n=24)	5 (20.8) [16,26,47-49]	5 (20.8) [61,88,89,115,116]	14 (58.3) [6,11,20,34,67-72,97,98,120,121]	0 (0)
Grand total (n=117)	42 (35.9)	38 (32.5)	36 (30.8)	1 (0.9)

[a]CSS: Cyberchondria Severity Scale.

[b]CSS-12: 12-item Cyberchondria Severity Scale.

In Table 2, of the 117 articles, the language of the instrument was unspecified in 24 (20.5%), as no explicit statement was provided. However, because the language can often be inferred from the national context in which the study was conducted, Table 3 reports scale use by language, incorporating both explicitly stated and inferred languages. In this revised analysis, studies implementing the traditional CSS in Turkish were most common (15/42, 36% articles), whereas among studies using the CSS-12, those implementing it in English (7/38, 18%) or Turkish predominated (7/38, 18%).

Table . Instrument use by language (languages inferred for ambiguous cases).

Language	CSS[a], n (%)	CSS-12[b], n (%)	Other, n (%)	Unknown, n (%)
Arabic (n=6)	0 (0)	4 (66.7) [27,80-82]	2 (33.3) [67,117]	0 (0)
Chinese (n=16)	2 (12.5) [35,74]	6 (37.5) [28,50,83,108-110]	8 (50) [6,34,68,71,72,90,120,121]	0 (0)
Croatian (n=6)	1 (16.7) [100]	0 (0)	5 (83.3) [7,18,91,92,118]	0 (0)
English (n=28)	9 (32.1) [8,9,12,13,21,22,36,47,48]	7 (25) [3,84,111-113,115,116]	11 (39.3) [10,11,20,32,62,69,70,93,97,98,119]	1 (3.6) [73]
French (n=1)	0 (0)	1 (100) [85]	0 (0)	0 (0)
German (n=3)	1 (33.3) [101]	0 (0)	2 (66.7) [19,63]	0 (0)
Indonesian (n=1)	0 (0)	0 (0)	1 (100) [64]	0 (0)
Italian (n=6)	3 (50) [14,37,38]	3 (50) [29,51,52]	0 (0)	0 (0)
Korean (n=1)	0 (0)	0 (0)	1 (100) [94]	0 (0)
Persian (n=5)	1 (20) [75]	4 (80) [17,30,53,61]	0 (0)	0 (0)
Polish (n=9)	7 (77.8) [15,23,24,39,102-104]	2 (22.2) [54,88]	0 (0)	0 (0)
Portuguese (n=1)	1 (100) [40]	0 (0)	0 (0)	0 (0)
Romanian (n=2)	2 (100) [16,26]	0 (0)	0 (0)	0 (0)
Russian (n=1)	0 (0)	1 (100) [55]	0 (0)	0 (0)
Serbian (n=2)	0 (0)	1 (50) [56]	1 (50) [95]	0 (0)
Spanish (n=2)	0 (0)	2 (100) [57,86]	0 (0)	0 (0)
Turkish (n=26)	15 (57.7) [25,41-46,49,76-79,105-107]	7 (26.9) [31,58-60,87,89,114]	4 (15.4) [33,65,66,96]	0 (0)
Urdu (n=1)	0 (0)	0 (0)	1 (100) [99]	0 (0)
Grand total (n=117)	42 (35.9)	38 (32.5)	36 (30.8)	1 (0.9)

[a]CSS: Cyberchondria Severity Scale.

[b]CSS-12: 12-item Cyberchondria Severity Scale.

As was the case in which languages were not inferred, a Fisher exact test found a significant association between the language in which an article implemented its cyberchondria measurement and the instrument used (P<.001); this association remained significant (P=.03) when articles using an instrument other than the CSS or CSS-12 were excluded.

Discussion

Adoption Trends

From the results in Table 1, it appears that the CSS-12 [3] had not completely replaced the CSS [1] in 2024. Given that there is no financial cost to switching instruments, it would be expected that the CSS-12 would completely replace the CSS over time if the 2 were perfect substitutes. This would be expected, as the CSS-12 is less time intensive to administer and is potentially less confusing for respondents due to its lack of reverse-keyed questions (eg, those measuring mistrust of medical professionals). The CSS-12's shorter length is potentially beneficial for both completion rates and the cost of administration. The main barriers to adoption of the CSS-12 in a study are researcher awareness and development of the study design after gaining awareness of the CSS-12. That said, the correlation between year and the proportion of cyberchondria studies using the CSS-12 achieved significance (ρ=0.89; *P*=.02), and it appears that there was a monotonic relationship trending toward greater use of the CSS-12 over time.

Measurement of the Mistrust of Medical Professionals

Given that the CSS-12 had been available for more than 4 years by the start of 2024, the fact that out of 23 studies, 8 (34.8%) used the original CSS in 2024 suggests that the CSS-12 may not be a perfect substitute. One key difference between the CSS and CSS-12 is that the CSS contains a subscale related to the mistrust of medical professionals, whereas the CSS-12 does not. Furthermore, this omission in the CSS-12 also makes it less suitable as an instrument for the study of the relationship between the mistrust of medical professionals and cyberchondria or other health issues, such as health anxiety [122].

Further research needs to be conducted to determine whether mistrust of medical professionals is a subconstruct related to, but distinct from, cyberchondria [2,13,38,123]. Concern over it being a distinct construct initially prompted its removal [3]. Some authors have opted to use the CSS without the reverse-keyed mistrust of medical professionals questions, citing concerns with the 5-factor structure of the CSS [8,24,36,47]. However, as the cyclical, reinforcing role of problematic digital information searches has been proposed to be a focal feature of cyberchondria presentations [124], barriers to accessing information from medical professionals constitute a concern of significant relevance. A lack of trust in health care providers broadly identifies a potential barrier to the access, use, and provision of care.

Measuring mistrust of medical professionals is relevant in public health and clinical care settings. Globally, most people do not trust medical professionals. The Wellcome Global Monitor 2020, a survey of more than 119,000 people residing in 113 countries, found that only 45% of people trust physicians and nurses in their country [125]. Measuring mistrust of medical professionals is increasingly important due to the erosion of trust that occurred during the COVID-19 pandemic. A repeated survey of Americans found that the proportion of adults who reported "a lot" of trust for physicians and hospitals declined from 71.5% in April 2020 to 40.1% in January 2024 [126]. These data suggest that the percentage of Americans with some doubts about the trustworthiness of medical professionals became the majority over this period. Furthermore, the study did not find signs that trust was rebounding. As mistrust of medical professionals becomes more common, it may be worth further exploring the nature of its association with cyberchondria, or its potential role as a control variable [127]. As these applications can only be performed with the original, long-form CSS, they provide a potential source of relevance for the measure going forward.

Moreover, measuring mistrust is important because cyberchondria can harm health care relationships between health care providers and patients in primary care settings [128,129] and may lead to "doctor shopping." Despite its impact on use, the degree to which mistrust impacts health care utilization has been underexplored [103]. Furthermore, health care providers may experience patients with cyberchondria as difficult to treat [128], which could lead to increased clinician burnout or stress. Outside of primary care settings, specifically within psychotherapy, strong care relationships are associated with positive outcomes [130]. This suggests that measuring and managing mistrust may alert health care providers to patients who may require additional communication or support. Additionally, across the reviewed literature, the importance of successful health care provider and patient communication is often referenced [49,51,54,73], and additional literature specifically mentions the importance of care alliances [131]. Consequently, identifying these patients may combat potential clinician burnout or stress and could arguably support successful care outcomes across medical and psychotherapeutic settings.

Infodemic-related concerns are also linked to cyberchondria [132] and are referenced in the reviewed literature [101,111,118]. This factor places strain on health systems [133] and may be of special relevance to the mistrust of medical professionals construct, as patients may encounter information online that contradicts their health care providers' recommendations. Digital literacy, for example, has been suggested as a supportive generalist cyberchondria intervention [116] and was included in the sole intervention identified in our review [14]. That said, higher digital literacy is also associated with higher cyberchondria scores, and the relationship may be mediated, moderated, or associated with other constructs [99,100,110,121].

Social Contagion in Instrument Selection

If an author uses an instrument while working on 1 study, or sees an instrument cited in a study written by someone within their professional or social network, they may be more likely to use it. Social contagion has been demonstrated in other clinical contexts. Specifically, it has been shown that there was social contagion in surgeons' adoption of perioperative advanced imaging when performing surgeries for the treatment of breast cancer. Patients treated by surgeons whose peers had the highest rates of imaging use were more likely to receive imaging than patients treated by surgeons whose peers had lower rates of use [134]. Likewise, social contagion may impact a researcher's desire to pursue a study on a topic such as cyberchondria.

In 2016, a German research group produced a 15-item version of the CSS in the German language [5]. It has been noted that some items loaded on different factors in the German implementation of the CSS than in the original version, creating a fundamental difference [5,14]. The German 15-item version of the CSS was translated into English and used by several India-based researchers [10,11,70]. This repeated use of a nonstandard version of the CSS may illustrate social contagion, especially because the modified instrument was reused in a different country and language than the one in which it originated. Additionally, both canonical versions of the scale are designed for English language use and therefore could be easier to deploy in a country that uses English as an official working language. Furthermore, the loading of items onto different factors in the German 15-item version of the CSS has the potential to reduce the comparability of studies based upon this implementation of the CSS versus other versions. Social contagion and ease of implementation may also explain why researchers using one language, such as Turkish, favor the CSS, while those using another, such as Chinese, favor the CSS-12. Further research is needed to assess the impact of social contagion on instrument selection.

Issues Related to Localization

There are both benefits and drawbacks to the localization of the CSS and CSS-12 into various languages. Providing patients with written materials in their native languages has been shown to improve comprehension [135]. However, translations of an instrument into a language may vary across researchers, leading to inconsistency in implementation, even when the same underlying instrument is used. For instance, some English medical terms have been shown to have multiple Arabic equivalents, potentially leading translations to differ [136]. Furthermore, the somatic features of depression have been shown to vary across cultures, suggesting that even standardized medical terms may be conceptualized and experienced differently by people in different contexts [137]. The developers of the Chinese CSS stated that cultural factors may influence both the presence of and responses to cyberchondria-like behaviors. Within a Chinese context, both linguistic and cultural factors influence instrument translation; "excessiveness" is a noteworthy example, as the authors explain that simply choosing to see a physician may be seen as excessive in China [74]. Finally, as the original CSS contains reverse-keyed questions and the CSS-12 does not, the CSS-12 may confer additional clarity or interpretation advantages when translated.

Multiple Hypothesis Testing

As the study used multiple hypothesis tests, it is possible that some statistically significant findings were false positives. The analysis included 6 Fisher exact tests and 1 Spearman rank correlation coefficient, for a total of 7 hypothesis tests. If the desired significance threshold is α=.05, then the Bonferroni correction implies that findings would remain significant only if the *P* value was <.007.

While Spearman rank correlation coefficient showed a statistically significant (*P*=.02) monotonic relationship between the year of publication and the proportion of studies using the CSS-12 before the Bonferroni correction was applied, the relationship was not statistically significant after considering the Bonferroni correction. The Fisher exact tests assessing the association between use of any cyberchondria scale and language of implementation, using the data in Tables 2 and 3, were all significant at the *P*<.001 level. These results therefore remained statistically significant after application of the Bonferroni correction. However, when the analyses were restricted to articles that used only the CSS or CSS-12, the Fisher exact tests for Tables 2 and 3 each yielded *P*=.03, which did not meet the Bonferroni-adjusted significance threshold.

Limitations

While this analysis captured the articles indexed by PubMed and PsycInfo, some relevant articles not included in these databases may have been missed. Likewise, there is often a body of gray literature consisting of unpublished manuscripts that are not publicly available due to their lack of significant findings, the direction of their findings, or abandonment by their authors. Therefore, while the findings do not necessarily represent all research conducted using the CSS, they do reflect the research accessible through 2 commonly used search tools, PubMed and PsycInfo.

Conclusions

This study examined how often the CSS and CSS-12 have been used in the literature, the languages in which they have been implemented, and the contexts in which each version may be preferable. From 2019 to 2024, both instruments continued to be used. Although the increasing adoption of the CSS-12 over time showed an unadjusted statistically significant monotonic trend (*P*=.02), this association did not remain significant after Bonferroni correction for multiple comparisons. The CSS-12 offers advantages such as brevity and the removal of reverse-keyed items, while the original CSS remains useful for studies that require the mistrust of medical professionals subscale. Researchers selecting an instrument should consider the benefits of shorter administration and improved clarity alongside the need to measure constructs unique to the full CSS, as well as the availability and quality of translations into the target population's language. Instrument choice should be guided by the study's objectives, the constructs of interest, and the cultural and linguistic context. Further research is needed to determine the interchangeability of adapted and translated versions with the original 33-item English CSS.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research or authorship of this article. Thomas Jefferson University provided partial support for the article processing fee.

Data Availability

Data sharing is not applicable to this paper as no datasets were generated or analyzed during this study.

Conflicts of Interest

ACP previously served on the Editorial Board of *JMIR Mental Health* and currently serves on the Editorial Board of the *Journal of Participatory Medicine*, a JMIR journal.

Multimedia Appendix 1
Process used to assign a scale to an article.
[PDF File, 67 KB - mental_v13i1e75003_app1.pdf]

Multimedia Appendix 2
Process used to assign a language to an article.
[PDF File, 67 KB - mental_v13i1e75003_app2.pdf]

References

1. McElroy E, Shevlin M. The development and initial validation of the cyberchondria severity scale (CSS). J Anxiety Disord 2014 Mar;28(2):259-265. [doi: 10.1016/j.janxdis.2013.12.007] [Medline: 24508033]
2. Fergus TA. The Cyberchondria Severity Scale (CSS): an examination of structure and relations with health anxiety in a community sample. J Anxiety Disord 2014 Aug;28(6):504-510. [doi: 10.1016/j.janxdis.2014.05.006] [Medline: 24956357]
3. McElroy E, Kearney M, Touhey J, Evans J, Cooke Y, Shevlin M. The CSS-12: development and validation of a short-form version of the cyberchondria severity scale. Cyberpsychol Behav Soc Netw 2019 May;22(5):330-335. [doi: 10.1089/cyber.2018.0624] [Medline: 31013440]
4. McMullan RD, Berle D, Arnáez S, Starcevic V. The relationships between health anxiety, online health information seeking, and cyberchondria: systematic review and meta-analysis. J Affect Disord 2019 Feb 15;245:270-278. [doi: 10.1016/j.jad.2018.11.037] [Medline: 30419526]
5. Barke A, Bleichhardt G, Rief W, Doering BK. The Cyberchondria Severity Scale (CSS): German validation and development of a short form. Int J Behav Med 2016 Oct;23(5):595-605. [doi: 10.1007/s12529-016-9549-8] [Medline: 26931780]
6. Hsu WC. Developing a cyberchondria severity scale to promote self-care among university students during COVID-19. Sci Rep 2024 Oct 10;14(1):23683. [doi: 10.1038/s41598-024-74829-z] [Medline: 39390121]
7. Jokić-Begić N, Mikac U, Čuržik D, Sangster Jokić C. The development and validation of the Short Cyberchondria Scale (SCS). J Psychopathol Behav Assess 2019 Dec;41(4):662-676. [doi: 10.1007/s10862-019-09744-z]
8. Gibler RC, Jastrowski Mano KE, O'Bryan EM, Beadel JR, McLeish AC. The role of pain catastrophizing in cyberchondria among emerging adults. Psychol Health Med 2019 Nov 26;24(10):1267-1276. [doi: 10.1080/13548506.2019.1605087]
9. Starcevic V, Baggio S, Berle D, Khazaal Y, Viswasam K. Cyberchondria and its relationships with related constructs: a network analysis. Psychiatr Q 2019 Sep;90(3):491-505. [doi: 10.1007/s11126-019-09640-5] [Medline: 31098922]
10. Dagar D, Kakodkar P, Shetiya SH. Evaluating the cyberchondria construct among computer engineering students in Pune (India) using cyberchondria severity scale (CSS-15). Indian J Occup Environ Med 2019;23(3):117-120. [doi: 10.4103/ijoem.IJOEM_217_19] [Medline: 31920260]
11. Makarla S, Gopichandran V, Tondare D. Prevalence and correlates of cyberchondria among professionals working in the information technology sector in Chennai, India: a cross-sectional study. J Postgrad Med 2019;65(2):87-92. [doi: 10.4103/jpgm.JPGM_293_18] [Medline: 31036778]
12. Akhtar M, Fatima T. Exploring cyberchondria and worry about health among individuals with no diagnosed medical condition. J Pak Med Assoc 2020 Jan;70(1):90-95. [doi: 10.5455/JPMA.8682] [Medline: 31954030]
13. Newby JM, McElroy E. The impact of internet-delivered cognitive behavioural therapy for health anxiety on cyberchondria. J Anxiety Disord 2020 Jan;69:102150. [doi: 10.1016/j.janxdis.2019.102150] [Medline: 31739276]
14. Marino C, Fergus TA, Vieno A, Bottesi G, Ghisi M, Spada MM. Testing the Italian version of the cyberchondria severity scale and a metacognitive model of cyberchondria. Clin Psychology and Psychoth 2020 Jul;27(4):581-596 [FREE Full text] [doi: 10.1002/cpp.2444]
15. Bajcar B, Babiak J. Neuroticism and cyberchondria: the mediating role of intolerance of uncertainty and defensive pessimism. Pers Individ Dif 2020 Aug;162:110006. [doi: 10.1016/j.paid.2020.110006]
16. Maftei A, Holman AC. Cyberchondria during the coronavirus pandemic: the effects of neuroticism and optimism. Front Psychol 2020;11:567345. [doi: 10.3389/fpsyg.2020.567345] [Medline: 33192848]
17. Seyed Hashemi SG, Hosseinnezhad S, Dini S, Griffiths MD, Lin CY, Pakpour AH. The mediating effect of the cyberchondria and anxiety sensitivity in the association between problematic internet use, metacognition beliefs, and fear of COVID-19 among Iranian online population. Heliyon 2020 Oct;6(10):e05135. [doi: 10.1016/j.heliyon.2020.e05135] [Medline: 33088933]
18. Jokic-Begic N, Lauri Korajlija A, Mikac U. Cyberchondria in the age of COVID-19. PLoS ONE 2020;15(12):e0243704. [doi: 10.1371/journal.pone.0243704] [Medline: 33332400]
19. Jungmann SM, Witthöft M. Health anxiety, cyberchondria, and coping in the current COVID-19 pandemic: which factors are related to coronavirus anxiety? J Anxiety Disord 2020 Jun;73:102239. [doi: 10.1016/j.janxdis.2020.102239] [Medline: 32502806]

20. Shailaja B, Shetty V, Chaudhury S, Thyloth M. Exploring cyberchondria and its associations in dental students amid COVID-19 infodemic. Ind Psychiatry J 2020;29(2):257-267. [doi: 10.4103/ipj.ipj_212_20]
21. Arsenakis S, Chatton A, Penzenstadler L, et al. Unveiling the relationships between cyberchondria and psychopathological symptoms. J Psychiatr Res 2021 Nov;143:254-261. [doi: 10.1016/j.jpsychires.2021.09.014] [Medline: 34509786]
22. Khazaal Y, Chatton A, Rochat L, et al. Compulsive health-related internet use and cyberchondria. Eur Addict Res 2021;27(1):58-66. [doi: 10.1159/000510922] [Medline: 33120393]
23. Bajcar B, Babiak J. Self-esteem and cyberchondria: the mediation effects of health anxiety and obsessive–compulsive symptoms in a community sample. Curr Psychol 2021 Jun;40(6):2820-2831. [doi: 10.1007/s12144-019-00216-x]
24. Oniszczenko W. Anxious temperament and cyberchondria as mediated by fear of COVID-19 infection: a cross-sectional study. PLoS ONE 2021;16(8):e0255750. [doi: 10.1371/journal.pone.0255750] [Medline: 34352033]
25. Köse S, Murat M. Examination of the relationship between smartphone addiction and cyberchondria in adolescents. Arch Psychiatr Nurs 2021 Dec;35(6):563-570. [doi: 10.1016/j.apnu.2021.08.009] [Medline: 34861946]
26. Maftei A, Holman A. Better once it's over, worse now: prospective moral behaviors after the coronavirus epidemic and cyberchondria. Psihologija 2021;54(2):193-205. [doi: 10.2298/PSI200603033M]
27. Rahme C, Akel M, Obeid S, Hallit S. Cyberchondria severity and quality of life among Lebanese adults: the mediating role of fear of COVID-19, depression, anxiety, stress and obsessive-compulsive behavior-a structural equation model approach. BMC Psychol 2021 Oct 29;9(1):169. [doi: 10.1186/s40359-021-00674-8] [Medline: 34715930]
28. Peng XQ, Chen Y, Zhang YC, et al. The status and influencing factors of cyberchondria during the COVID-19 epidemic. A cross-sectional study in Nanyang City of China. Front Psychol 2021;12:712703. [doi: 10.3389/fpsyg.2021.712703] [Medline: 34858254]
29. Vismara M, Vitella D, Biolcati R, et al. The impact of COVID-19 pandemic on searching for health-related information and cyberchondria on the general population in Italy. Front Psychiatry 2021;12:754870. [doi: 10.3389/fpsyt.2021.754870] [Medline: 34712159]
30. Wu X, Nazari N, Griffiths MD. Using fear and anxiety related to COVID-19 to predict cyberchondria: cross-sectional survey study. J Med Internet Res 2021 Jun 9;23(6):e26285. [doi: 10.2196/26285] [Medline: 34014833]
31. Yam FC, Korkmaz O, Griffiths MD. The association between fear of COVID-19 and smartphone addiction among individuals: the mediating and moderating role of cyberchondria severity. Curr Psychol 2023;42(3):2377-2390. [doi: 10.1007/s12144-021-02324-z] [Medline: 34690474]
32. Bala R, Srivastava A, Ningthoujam GD, Potsangbam T, Oinam A, Anal CL. An observational study in Manipur state, India on preventive behavior influenced by social media during the COVID-19 pandemic mediated by cyberchondria and information overload. J Prev Med Public Health 2021 Jan;54(1):22-30. [doi: 10.3961/jpmph.20.465] [Medline: 33618496]
33. Durak Batıgün A, Şenkal Ertürk İ, Gör N, Kömürcü Akik B. The pathways from distress tolerance to cyberchondria: a multiple-group path model of young and middle adulthood samples. Curr Psychol 2021;40(11):5718-5726. [doi: 10.1007/s12144-020-01038-y] [Medline: 32921966]
34. Han L, Zhan Y, Li W, Xu Y, Xu Y, Zhao J. Associations between the perceived severity of the COVID-19 pandemic, cyberchondria, depression, anxiety, stress, and lockdown experience: cross-sectional survey study. JMIR Public Health Surveill 2021 Sep 16;7(9):e31052. [doi: 10.2196/31052] [Medline: 34478402]
35. Zhou Y, Dai L, Deng Y, Zeng H, Yang L. The moderating effect of alexithymia on the relationship between stress and cyberchondria. Front Psychiatry 2022;13:1043521. [doi: 10.3389/fpsyt.2022.1043521]
36. Airoldi S, Kolubinski DC, Nikčević AV, Spada MM. The relative contribution of health cognitions and metacognitions about health anxiety to cyberchondria: a prospective study. J Clin Psychol 2022 May;78(5):809-820. [doi: 10.1002/jclp.23252] [Medline: 34559886]
37. Bottesi G, Marino C, Vieno A, Ghisi M, Spada MM. Psychological distress in the context of the COVID-19 pandemic: the joint contribution of intolerance of uncertainty and cyberchondria. Psychol Health 2022 Nov 2;37(11):1396-1413. [doi: 10.1080/08870446.2021.1952584]
38. Vismara M, Benatti B, Ferrara L, et al. A preliminary investigation of cyberchondria and its correlates in a clinical sample of patients with obsessive–compulsive disorder, anxiety and depressive disorders attending a tertiary psychiatric clinic. Int J Psychiatry Clin Pract 2022 Jun;26(2):111-122. [doi: 10.1080/13651501.2021.1927107] [Medline: 34032529]
39. Ciułkowicz M, Misiak B, Szcześniak D, Grzebieluch J, Maciaszek J, Rymaszewska J. The portrait of cyberchondria-a cross-sectional online study on factors related to health anxiety and cyberchondria in Polish population during SARS-CoV-2 pandemic. Int J Environ Res Public Health 2022 Apr 5;19(7):4347. [doi: 10.3390/ijerph19074347] [Medline: 35410027]
40. Serra-Negra JM, Paiva SM, Baptista AS, Cruz AJ, Pinho T, Abreu MH. Cyberchondria and associated factors among Brazilian and Portuguese dentists. Acta Odontol Latinoam 2022 Apr 30;35(1):45-50. [doi: 10.54589/aol.35/1/45] [Medline: 35700541]
41. Demirtas Z, Emiral GO, Caliskan S, et al. Evaluation of relationship between cyberchondria and obsessive beliefs in adults. P R Health Sci J 2022 Dec;41(4):233-238 [FREE Full text] [Medline: 36516210]
42. Karakaş N, Tekin Ç, Bentli R, Demir E. Cyberchondria, COVID-19 phobia, and well-being: a relational study on teachers. Med Lav 2022 Jun 28;113(3):e2022027. [doi: 10.23749/mdl.v113i3.12661] [Medline: 35766648]

43. Kılıçaslan AK, Yıldız S, Gür C C, Uğur K. Cyberchondria and health anxiety in patients with fibromyalgia. Arch Psych Psych 2022;24(4):16-25. [doi: 10.12740/APP/150478]
44. Özkan O, Sungur C, Özer Ö. Investigation of cyberchondria level and digital literacy on women in Turkey. J Hum Behav Soc Environ 2022 Aug 18;32(6):768-780. [doi: 10.1080/10911359.2021.1962776]
45. Sezer Ö, Başoğlu MA, Dağdeviren HN. An examination of cyberchondria's relationship with trait anxiety and psychological well-being in women of reproductive age: a cross-sectional study. Medicine (Baltimore) 2022 Nov 18;101(46):e31503. [doi: 10.1097/MD.0000000000031503] [Medline: 36401487]
46. Uysal Toraman A, Kalkim A, Korkmaz EK. Coronavirus anxiety and cyberchondria among teachers during the COVID-19 pandemic: an online survey. Curr Psychol 2024 Apr;43(14):13219-13225. [doi: 10.1007/s12144-022-03382-7]
47. Nadeem F, Malik NI, Atta M, et al. Relationship between health-anxiety and cyberchondria: role of metacognitive beliefs. J Clin Med 2022 May 5;11(9):2590. [doi: 10.3390/jcm11092590] [Medline: 35566713]
48. Sohail M, Zafar N. Fear of COVID-19 and stress in university students: mediating role of cyberchondria and moderating role of creative coping and social supports. J Pak Med Assoc 2022 Aug 1;72(8):1564-1571. [doi: 10.47391/JPMA.4350]
49. Turhan Cakir A. Cyberchondria levels in women with human papilloma virus. J Obstet Gynaecol Res 2022 Oct;48(10):2610-2614. [doi: 10.1111/jog.15354] [Medline: 35801694]
50. Liu S, Yang H, Cheng M, Miao T. Family dysfunction and cyberchondria among Chinese adolescents: a moderated mediation model. Int J Environ Res Public Health 2022 Aug 7;19(15):9716. [doi: 10.3390/ijerph19159716] [Medline: 35955070]
51. Ambrosini F, Truzoli R, Vismara M, Vitella D, Biolcati R. The effect of cyberchondria on anxiety, depression and quality of life during COVID-19: the mediational role of obsessive-compulsive symptoms and Internet addiction. Heliyon 2022 May;8(5):e09437. [doi: 10.1016/j.heliyon.2022.e09437] [Medline: 35600442]
52. Santoro G, Starcevic V, Scalone A, Cavallo J, Musetti A, Schimmenti A. The doctor is in(ternet): the mediating role of health anxiety in the relationship between somatic symptoms and cyberchondria. J Pers Med 2022 Sep 12;12(9):1490. [doi: 10.3390/jpm12091490] [Medline: 36143275]
53. Ahorsu DK, Lin CY, Alimoradi Z, et al. Cyberchondria, fear of COVID-19, and risk perception mediate the association between problematic social media use and intention to get a COVID-19 vaccine. Vaccines (Basel) 2022 Jan 14;10(1):122. [doi: 10.3390/vaccines10010122] [Medline: 35062783]
54. Błachnio A, Przepiórka A, Kot P, Cudo A, Steuden S. The role of emotional functioning in the relationship between health anxiety and cyberchondria. Curr Psychol 2023 Dec;42(35):31240-31250. [doi: 10.1007/s12144-022-04126-3]
55. Zolotareva A. Cyberchondria, but not preventive behavior, mediates the relationship between fear of COVID-19 and somatic burden: evidence from Russia. Front Psychiatry 2022;13:1018659. [doi: 10.3389/fpsyt.2022.1018659] [Medline: 36226097]
56. Vujić A, Dinić BM, Jokić-begić N. Cyberchondria and questionable health practices: the mediation role of conspiracy mentality. SP 2022 Mar 16;64(1):104-117. [doi: 10.31577/sp.2022.01.842]
57. Arnáez S, García-Soriano G, Castro J, Berle D, Starcevic V. The Spanish version of the short form of the Cyberchondria Severity Scale (CSS-12): testing the factor structure and measurement invariance across genders. Curr Psychol 2023 Aug;42(24):20686-20695. [doi: 10.1007/s12144-022-03170-3]
58. Boysan M, Eşkisu M, Çam Z. Relationships between fear of COVID-19, cyberchondria, intolerance of uncertainty, and obsessional probabilistic inferences: a structural equation model. Scand J Psychol 2022 Oct;63(5):439-448. [doi: 10.1111/sjop.12822] [Medline: 35430750]
59. Varer Akpinar C, Mandiracioglu A, Ozvurmaz S, Kurt F, Koc N. Cyberchondria and COVID-19 anxiety and internet addiction among nursing students. Curr Psychol 2023;42(3):2406-2414. [doi: 10.1007/s12144-022-04057-z] [Medline: 36468163]
60. Yalçın İ, Boysan M, Eşkisu M, Çam Z. Health anxiety model of cyberchondria, fears, obsessions, sleep quality, and negative affect during COVID-19. Curr Psychol 2024 Mar;43(9):8502-8519. [doi: 10.1007/s12144-022-02987-2]
61. Daniali M, Eskandari E. Predicting coronavirus anxiety based on resilience, cognitive emotion regulation strategies, and cyberchondria. Advance Cogn Sci 2022 Jan 10;23(4):61-71 [FREE Full text] [doi: 10.30514/icss.23.4.61]
62. Mrayyan MT, AL‐Atiyyat N, Abu Khait A, Al‐Rawashdeh S, Algunmeeyn A, Abunab HY. Does cyberchondria predict Internet addiction among students during the COVID‐19 pandemic? A web‐based survey study. Nurs Forum 2022 May;57(3):337-343. [doi: 10.1111/nuf.12682]
63. Nicolai J, Moshagen M, Schillings K, Erdfelder E. The role of base-rate neglect in cyberchondria and health anxiety. J Anxiety Disord 2022 Oct;91:102609. [doi: 10.1016/j.janxdis.2022.102609] [Medline: 35963146]
64. Honora A, Wang KY, Chih WH. How does information overload about COVID-19 vaccines influence individuals' vaccination intentions? The roles of cyberchondria, perceived risk, and vaccine skepticism. Comput Human Behav 2022 May;130:107176. [doi: 10.1016/j.chb.2021.107176] [Medline: 35013641]
65. Durmuş A, Deniz S, Akbolat M, Çimen M. Does cyberchondria mediate the effect of COVID-19 fear on the stress? Soc Work Public Health 2022 May 19;37(4):356-369. [doi: 10.1080/19371918.2021.2014013] [Medline: 35100946]
66. Kurcer MA, Erdogan Z, Cakir Kardes V. The effect of the COVID‐19 pandemic on health anxiety and cyberchondria levels of university students. Perspect Psychiatric Care 2022 Jan;58(1):132-140. [doi: 10.1111/ppc.12850]

67. Abu Khait A, Mrayyan MT, Al-Rjoub S, Rababa M, Al-Rawashdeh S. Cyberchondria, anxiety sensitivity, hypochondria, and internet addiction: implications for mental health professionals. Curr Psychol 2023 Nov;42(31):27141-27152. [doi: 10.1007/s12144-022-03815-3]
68. Li J. Impact of metaverse cultural communication on the mental health of international students in China: highlighting effects of healthcare anxiety and cyberchondria. Am J Health Behav 2022 Dec 30;46(6):809-820. [doi: 10.5993/AJHB.46.6.21] [Medline: 36721290]
69. Patanapu SK, Sreeja CS, Veeraboina N, Reddy KV, Voruganti S, Anusha P. Prevalence and effect of cyberchondria on academic performance among undergraduate dental students: an institutional based study. Ind Psychiatry J 2022;31(2):228-234. [doi: 10.4103/ipj.ipj_272_21] [Medline: 36419676]
70. Pawar P, Kamat A, Salimath G, Jacob KR, Kamath R. Prevalence of cyberchondria among outpatients with metabolic syndrome in a tertiary care hospital in southern India. ScientificWorldJournal 2022;2022(1):3211501. [doi: 10.1155/2022/3211501] [Medline: 36199439]
71. Yuan W. Identifying the effect of digital healthcare products in metaverse on mental health: studying the interaction of cyberchondria and technophobia. Am J Health Behav 2022 Dec 30;46(6):729-739. [doi: 10.5993/AJHB.46.6.15] [Medline: 36721275]
72. Zheng H, Jiang S. Linking the pathway from exposure to online vaccine information to cyberchondria during the COVID-19 pandemic: a moderated mediation model. Cyberpsychol Behav Soc Netw 2022 Oct;25(10):625-633. [doi: 10.1089/cyber.2022.0045] [Medline: 36037024]
73. Afrin R, Prybutok G. Insights into the antecedents of cyberchondria: a perspective from the USA. Health Promot Int 2022 Aug 1;37(4):daac108. [doi: 10.1093/heapro/daac108] [Medline: 36047641]
74. Wang D, Sun L, Shao Y, Zhang X, Maguire P, Hu Y. Research and evaluation of a cyberchondria severity scale in a Chinese context. Psychol Res Behav Manag 2023;16:4417-4429. [doi: 10.2147/PRBM.S431470] [Medline: 37936970]
75. Nasiri M, Mohammadkhani S, Akbari M, Alilou MM. The structural model of cyberchondria based on personality traits, health-related metacognition, cognitive bias, and emotion dysregulation. Front Psychiatry 2022;13:960055. [doi: 10.3389/fpsyt.2022.960055] [Medline: 36699479]
76. Aydın Kartal Y, Kaya L, Özcan H. Investigation of the relationship between depression, cyberchondria levels and the quality of life of female students during the COVID-19 pandemic. Women Health 2023 Sep 14;63(8):669-680. [doi: 10.1080/03630242.2023.2255312]
77. Sayar SE, Demet Ust Tasgin Z, Gundogdu G. The relationship between fear of COVID-19 and levels of cyberchondria and evaluation of affecting factors. Psychiat Danub 2023 Oct 23;35(3):418-429. [doi: 10.24869/psyd.2023.418]
78. Özer Ö, Özmen S, Özkan O. Investigation of the effect of cyberchondria behavior on e-health literacy in healthcare workers. Hosp Top 2023;101(2):94-102. [doi: 10.1080/00185868.2021.1969873] [Medline: 34461810]
79. Üzüm Ö, Ince G, Eliaçık K, Kanık A, Elmalı F, Helvacı M. Investigating the potential connection between cyberchondria and vaccine hesitancy in high school students. Cureus 2023 Jan;15(1):e34218. [doi: 10.7759/cureus.34218] [Medline: 36852372]
80. El-Zayat A, Namnkani SA, Alshareef NA, Mustfa MM, Eminaga NS, Algarni GA. Cyberchondria and its association with smartphone addiction and electronic health literacy among a Saudi population. Saudi J Med Med Sci 2023;11(2):162-168. [doi: 10.4103/sjmms.sjmms_491_22] [Medline: 37252023]
81. Hallit S, Rogoza R, Abi Semaan C, Azzi V, Sawma T, Obeid S. Validation of the Arabic version of the cyberchondria severity scale 12 items (CSS-12-Ar) among a sample of Lebanese adults. BMC Psychiatry 2023 Aug 23;23(1):618. [doi: 10.1186/s12888-023-05123-x]
82. Tarabay C, Bitar Z, Akel M, Hallit S, Obeid S, Soufia M. Cyberchondria severity and quality of life among Lebanese adults: the moderating effect of emotions. Prim Care Companion CNS Disord 2023 Apr 27;25(2):46791. [doi: 10.4088/PCC.22m03252] [Medline: 37115149]
83. Zhu X, Zheng T, Ding L, Zhang X. Exploring associations between eHealth literacy, cyberchondria, online health information seeking and sleep quality among university students: a cross-section study. Heliyon 2023 Jun;9(6):e17521. [doi: 10.1016/j.heliyon.2023.e17521]
84. Sabir S, Naqvi I. Prevalence of cyberchondria among university students: an emerging challenge of the 21st century. J Pak Med Assoc 2023 Aug;73(8):1634-1639. [doi: 10.47391/JPMA.7771] [Medline: 37697754]
85. Infanti A, Starcevic V, Schimmenti A, et al. Predictors of cyberchondria during the COVID-19 pandemic: cross-sectional study using supervised machine learning. JMIR Form Res 2023 Apr 25;7(1):e42206. [doi: 10.2196/42206] [Medline: 36947575]
86. Robles-Mariños R, Alvarado GF, Maguiña JL, Bazo-Alvarez JC. The short-form of the Cyberchondria Severity Scale (CSS-12): adaptation and validation of the Spanish version in young Peruvian students. PLoS ONE 2023;18(10):e0292459. [doi: 10.1371/journal.pone.0292459] [Medline: 37796833]
87. Yıldız M, Demirhan A, Gökçay G, Polat F. The relationship between cyberchondria levels, attitudes towards menopause and menopausal complaints of women in the climacteric period: analysis with data mining. Womens Stud Int Forum 2023 May;98:102701. [doi: 10.1016/j.wsif.2023.102701]

88. Błachnio A, Przepiórka A, Kot P, Cudo A, McElroy E. The mediating role of rumination between stress appraisal and cyberchondria. Acta Psychol (Amst) 2023 Aug;238:103946. [doi: 10.1016/j.actpsy.2023.103946] [Medline: 37499622]
89. Özkent MS, Kılınç MT, Hamarat MB, et al. Digitalization and urological diseases: severity of cyberchondria and level of health anxiety in patients visiting outpatient urology clinics. Cyberpsychol Behav Soc Netw 2023 Jan;26(1):28-34. [doi: 10.1089/cyber.2022.0089] [Medline: 36454182]
90. Liu Y, Peng W, Cao M, Zhang S, Peng J, Zhou Z. Cyberchondria and Chinese adolescent mental health in the age of COVID-19 pandemic. Cyberpsychol Behav Soc Netw 2023 Aug;26(8):631-639. [doi: 10.1089/cyber.2022.0319] [Medline: 37406285]
91. Bagaric B, Martincevic M, Vranic A. What is remembered?: the recall of health-related information in cyberchondria and health anxiety. Psihologija 2023;56(2):205-221. [doi: 10.2298/PSI220127019B]
92. Šoštarić M, Mikac U, Jokić-Begić N. Understanding cyberchondria in pregnant women: longitudinal assessment of risk factors, triggers, and outcomes. J Psychosom Obstet Gynecol 2023 Dec 31;44(1):2265050. [doi: 10.1080/0167482X.2023.2265050]
93. Mrayyan MT, Alkhawaldeh JM, Alfayoumi I, et al. COVID-19 era-related e-learning: a cross-sectional web-scale study of cyberchondria, internet addiction and anxiety-related symptomatology among university nursing students. BMJ Open 2023 Aug 9;13(8):e071971. [doi: 10.1136/bmjopen-2023-071971] [Medline: 37558438]
94. Jeong GC, Lee K, Jin Y. Effects of the fear of COVID-19 and efficacy of coping behavior for infectious diseases after the end of COVID-19: moderating effects of cyberchondria and eHealth literacy. Behav Sci (Basel) 2023 Aug 8;13(8):663. [doi: 10.3390/bs13080663] [Medline: 37622803]
95. Vujić A, Volarov M, Latas M, Demetrovics Z, Kiraly O, Szabo A. Are cyberchondria and intolerance of uncertainty related to smartphone addiction? Int J Ment Health Addiction 2024 Dec;22(6):3361-3379. [doi: 10.1007/s11469-023-01054-6]
96. Ustuner Top F, Çevik C, Bora Güneş N. The relation between digital literacy, cyberchondria, and parents’ attitudes to childhood vaccines. J Pediatr Nurs 2023 May;70:12-19. [doi: 10.1016/j.pedn.2023.01.006]
97. Abikoye GE, Lawal AM. Prevalence and psychosocial predictors of cyberchondria in Nigeria during the COVID-19 pandemic. Int J Cyber Behav Psychol Learn 2023 Jan;13(1):1-12 [FREE Full text] [doi: 10.4018/IJCBPL.324088]
98. Satyarup D, Panda S, Nagarajappa R, Mohapatra U. Cyberchondria among information technology professionals of Bhubaneswar by using cyberchondria severity scale (CSS-15). Rocz Panstw Zakl Hig 2023;74(1):83-91. [doi: 10.32394/rpzh.2023.0241] [Medline: 37013889]
99. Shahani R, Asmi F, Ma J, et al. How cyberchondria and decision self-efficacy shapes the acceptability of COVID-19 vaccine: a gender-based comparison. Digit Health 2023;9:20552076231185430. [doi: 10.1177/20552076231185430] [Medline: 37744744]
100. Staraj Bajcic T, Sorta-Bilajac Turina I, Lucijanic M, Sinozic T, Vuckovic M, Bazdaric K. Cyberchondria, health literacy, and perception of risk in Croatian patients with risk of sexually transmitted infections and HIV-A cross-sectional study. Epidemiologia (Basel) 2024 Aug 22;5(3):525-538. [doi: 10.3390/epidemiologia5030036] [Medline: 39311353]
101. Jungmann SM, Gropalis M, Schenkel SK, Witthöft M. Is cyberchondria specific to hypochondriasis? J Anxiety Disord 2024 Mar;102:102798. [doi: 10.1016/j.janxdis.2023.102798] [Medline: 38128287]
102. Jędrzejewska AB, Nowicki GJ, Rudnicka-Drożak E, Panasiuk L, Ślusarska BJ. Association between cyberchondria and the use of complementary and alternative medicine (CAM) - a cross-sectional study. Ann Agric Environ Med 2024 Mar 25;31(1):87-93. [doi: 10.26444/aaem/178503] [Medline: 38549481]
103. Kobryn M, Duplaga M. Does health literacy protect against cyberchondria: a cross-sectional study? Telemed J E Health 2024 Apr;30(4):e1089-e1100. [doi: 10.1089/tmj.2023.0425] [Medline: 38016126]
104. Kobryn M, Duplaga M. Cyberchondria severity and utilization of health services in Polish society: a cross-sectional study. BMC Public Health 2024 Mar 27;24(1):902. [doi: 10.1186/s12889-024-18399-9] [Medline: 38539164]
105. Atsizata M, Sögüt SC. The relationship between orthorexia nervosa and cyberchondria levels in nurses: a cross-sectional study. Arch Psychiatr Nurs 2024 Feb;48:30-35. [doi: 10.1016/j.apnu.2024.01.008] [Medline: 38453279]
106. Bahadir O, Dundar C. The impact of online health information source preference on intolerance to uncertainty and cyberchondria in a youthful generation. Indian J Psychiatry 2024 Apr;66(4):360-366. [doi: 10.4103/indianjpsychiatry.indianjpsychiatry_715_23] [Medline: 38778859]
107. Topkara Sucu S, Kolomuç Gayretli T, Küçükkayıkçı AS, et al. Cyberchondria levels in adolescent patients with polycystic ovary syndrome in the digital age. J Pediatr Adolesc Gynecol 2024 Dec;37(6):569-573. [doi: 10.1016/j.jpag.2024.08.006] [Medline: 39168278]
108. Fang J, Qiu C, Sun Z, et al. A national survey of pandemic fear and cyberchondria after ending zero-COVID policy: the chain mediating role of alexithymia and psychological distress. Compr Psychiatry 2024 Aug;133:152505. [doi: 10.1016/j.comppsych.2024.152505] [Medline: 38852302]
109. Li Y, Li J, Zhou C, et al. Unraveling the relationships among pandemic fear, cyberchondria, and alexithymia after China’s exit from the zero-COVID policy: insights from a multi-center network analysis. Front Psychiatry 2024 Nov 14;15:1489961. [doi: 10.3389/fpsyt.2024.1489961]

110. Xu RH, Chen C. Moderating effect of coping strategies on the association between the infodemic-driven overuse of health care services and cyberchondria and anxiety: partial least squares structural equation modeling study. J Med Internet Res 2024 Apr 9;26:e53417. [doi: 10.2196/53417] [Medline: 38593427]
111. Agrawal V, Khulbe Y, Singh A, Kar SK. The digital health dilemma: exploring cyberchondria, well-being, and smartphone addiction in medical and non-medical undergraduates. Indian J Psychiatry 2024 Mar;66(3):256-262. [doi: 10.4103/indianjpsychiatry.indianjpsychiatry_570_23] [Medline: 39100122]
112. Ali SS, Hendawi NE, El-Ashry AM, Mohammed MS. The relationship between cyberchondria and health literacy among first-year nursing students: the mediating effect of health anxiety. BMC Nurs 2024 Oct 22;23(1):776. [doi: 10.1186/s12912-024-02396-9] [Medline: 39434055]
113. El-Zoghby SM, Zaghloul NM, Tawfik AM, Elsherbiny NM, Shehata SA, Soltan EM. Cyberchondria and smartphone addiction: a correlation survey among undergraduate medical students in Egypt. J Egypt Public Health Assoc 2024 Apr 3;99(1):7. [doi: 10.1186/s42506-024-00154-y] [Medline: 38565743]
114. Eşkisu M, Çam Z, Boysan M. Health-related cognitions and metacognitions indirectly contribute to the relationships between impulsivity, fear of COVID-19, and cyberchondria. J Rat Emo Cognitive Behav Ther 2024 Mar;42(1):110-132. [doi: 10.1007/s10942-022-00495-7]
115. Fang S, Mushtaque I. The moderating role of health literacy and health promoting behavior in the relationship among health anxiety, emotional regulation, and cyberchondria. Psychol Res Behav Manag 2024;17:51-62. [doi: 10.2147/PRBM.S446448] [Medline: 38196775]
116. Sansakorn P, Mushtaque I, Awais-E-Yazdan M, Dost MK. The relationship between cyberchondria and health anxiety and the moderating role of health literacy among the Pakistani public. Int J Environ Res Public Health 2024 Sep 2;21(9):1168. [doi: 10.3390/ijerph21091168] [Medline: 39338051]
117. Bin Abdulrahman KA, AL Musfir SK, Alforaih AS, Alshehri AM, Aldossari AK, Dawood FDB. The prevalence of cyberchondria and the impact of skepticism on medical decisions among Imam Mohammed Ibn Saud Islamic University students, Riyadh, Saudi Arabia. J Family Med Prim Care 2024;13(11):5334-5340. [doi: 10.4103/jfmpc.jfmpc_640_24]
118. Šoštarić M, Jokić-Begić N, Vukušić Mijačika M. Can't stop, won't stop - understanding anxiety's role in cyberchondria among pregnant women. Women Health 2024 Feb 7;64(2):185-194. [doi: 10.1080/03630242.2024.2308525] [Medline: 38258443]
119. Mrayyan MT, Abu Khait A, Al-Mrayat Y, et al. Anxiety sensitivity moderates the relationship between internet addiction and cyberchondria among nurses. J Health Psychol 2025 Sep;30(11):3125-3136. [doi: 10.1177/13591053241249634]
120. Guo Y, Wang Y, Li Y. Online health information seeking, health anxiety and cyberchondria among men who engage in sexual risk taking: the mediating role of medical consultation about HIV / AIDS. Community Applied Soc Psy 2024 Jul;34(4):e2845 [FREE Full text] [doi: 10.1002/casp.2845]
121. Zhang X, Zheng H, Zeng Y, Zou J, Zhao L. Exploring how health-related advertising interference contributes to the development of cyberchondria: a stressor-strain-outcome approach. Digit Health 2024;10:20552076241233138. [doi: 10.1177/20552076241233138] [Medline: 38384368]
122. Salkovskis PM, Warwick H. Making sense of hypochondriasis: a cognitive theory of health anxiety. In: Asmundson G, Taylor S, Cox BJ, editors. Health Anxiety: Clinical and Research Perspectives on Hypochondriasis and Related Conditions: Wiley; 2001:46-64.
123. Norr AM, Allan NP, Boffa JW, Raines AM, Schmidt NB. Validation of the Cyberchondria Severity Scale (CSS): replication and extension with bifactor modeling. J Anxiety Disord 2015 Apr;31:58-64. [doi: 10.1016/j.janxdis.2015.02.001] [Medline: 25734759]
124. Starcevic V, Berle D. Cyberchondria: towards a better understanding of excessive health-related Internet use. Expert Rev Neurother 2013 Feb;13(2):205-213. [doi: 10.1586/ern.12.162] [Medline: 23368807]
125. Wellcome Trust, Gallup. Wellcome Global Monitor 2020: how COVID-19 affected people's lives and their views about science.: Wellcome Trust; 2021. URL: https://cms.wellcome.org/sites/default/files/2021-11/Wellcome-Global-Monitor-Covid.pdf [accessed 2025-08-13]
126. Perlis RH, Ognyanova K, Uslu A, et al. Trust in physicians and hospitals during the COVID-19 pandemic in a 50-state survey of US adults. JAMA Netw Open 2024 Jul 1;7(7):e2424984. [doi: 10.1001/jamanetworkopen.2024.24984] [Medline: 39083270]
127. Bajcar B, Babiak J, Olchowska-Kotala A. Cyberchondria and its measurement. The Polish adaptation and psychometric properties of the Cyberchondria Severity Scale CSS-PL. Psychiatr Pol 2019;53(1):49-60. [doi: 10.12740/PP/81799]
128. Wangler J, Jansky M. General practitioners' challenges and strategies in dealing with Internet-related health anxieties—results of a qualitative study among primary care physicians in Germany. Wien Med Wochenschr 2020 Oct;170(13-14):329-339. [doi: 10.1007/s10354-020-00777-8]
129. Wangler J, Jansky M. Online enquiries and health concerns – a survey of German general practitioners regarding experiences and strategies in patient care. J Public Health (Berl) 2024 Jul;32(7):1243-1249. [doi: 10.1007/s10389-023-01909-1]
130. Gelso CJ, Kivlighan DM, Markin RD. The real relationship and its role in psychotherapy outcome: a meta-analysis. Psychotherapy (Chic) 2018 Dec;55(4):434-444. [doi: 10.1037/pst0000183] [Medline: 30335456]

131. Vismara M, Caricasole V, Starcevic V, et al. Is cyberchondria a new transdiagnostic digital compulsive syndrome? A systematic review of the evidence. Compr Psychiatry 2020 May;99:152167. [doi: 10.1016/j.comppsych.2020.152167] [Medline: 32146315]
132. Laato S, Islam AKMN, Islam MN, Whelan E. What drives unverified information sharing and cyberchondria during the COVID-19 pandemic? Eur J Inf Syst 2020 May 3;29(3):288-305. [doi: 10.1080/0960085X.2020.1770632]
133. García-Saisó S, Marti M, Brooks I, et al. The COVID-19 Infodemic. Rev Panam Salud Publica 2021;45:e56. [doi: 10.26633/RPSP.2021.56] [Medline: 34234820]
134. Pollack CE, Soulos PR, Herrin J, et al. The impact of social contagion on physician adoption of advanced imaging tests in breast cancer. J Natl Cancer Inst 2017 Aug 1;109(8):djw330. [doi: 10.1093/jnci/djw330] [Medline: 28376191]
135. Perera KYS, Ranasinghe P, Adikari AMMC, Balagobi B, Constantine GR, Jayasinghe S. Medium of language in discharge summaries: would the use of native language improve patients' knowledge of their illness and medications? J Health Commun 2012;17(2):141-148. [doi: 10.1080/10810730.2011.585926] [Medline: 22112212]
136. Al-Jarf RS. Multiple Arabic equivalents to English medical terms. Int Linguist Res 2018;1(1):102. [doi: 10.30560/ilr.v1n1p102]
137. Chentsova-Dutton YE, Tsai JL. Understanding depression across cultures. In: Friedman ES, Anderson IM, editors. Handbook of Depression, 2nd edition: Guilford Press; 2009:363-385.

Abbreviations

CSS: Cyberchondria Severity Scale
CSS-12: 12-item Cyberchondria Severity Scale
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by J Torous; submitted 26.Mar.2025; peer-reviewed by R Ghosh, R Afrin; revised version received 29.Nov.2025; accepted 13.Dec.2025; published 05.Jan.2026.

Please cite as:
Powell AC, Calderon-Smith C
Trends in the Implementation of the Cyberchondria Severity Scale: Bibliometric Analysis
JMIR Ment Health 2026;13:e75003
URL: https://mental.jmir.org/2026/1/e75003
doi:10.2196/75003

Detecting Pediatric Emergency Service Use for Suicide and Self-Harm: Multimodal Analysis of 3828 Encounters

Juliet Beni Edgcomb[1], MD, PhD; Angshuman Saha[2], PhD; Alexandra Klomhaus[2], PhD; Elyse Tascione[1], MA; Chrislie G Ponce[1], BA; Joshua J Lee[2], BS; Theona Tacorda[3], MS; Bonnie T Zima[1], MD, MPH

[1]Mental Health Informatics and Data Science (MINDS) Hub, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, 760 Westwood Plz, Los Angeles, CA, United States

[2]Department of Medicine Statistics Core, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States

[3]Biomedical Informatics Program, Clinical and Translational Science Institute, University of California, Los Angeles, Los Angeles, CA, United States

Corresponding Author:
Juliet Beni Edgcomb, MD, PhD
Mental Health Informatics and Data Science (MINDS) Hub, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, 760 Westwood Plz, Los Angeles, CA, United States

Abstract

Background: Suicide is the second-leading cause of US childhood mortality after 9 years of age. The accurate measurement of pediatric emergency service use for self-injurious thoughts and behaviors (SITB) remains challenging, as diagnostic codes undercount children. This measurement gap impedes public health and prevention efforts. Current research has not established which combination of electronic health record data elements achieves both high detection accuracy and consistent performance across youth populations.

Objective: This study aims to (1) compare the detection accuracy of electronic health record−based methods for identifying SITB-related pediatric emergency department (ED) visits: basic structured data (*International Classification of Diseases Version 10, Clinical Modification* codes, chief concern), comprehensive structured data, clinical note text with natural language processing, and hybrid approaches combining structured data with notes; and (2) for each method, measure variability in detection by youth demographics and underlying mental health diagnosis.

Methods: Multiple human experts reviewed clinical records of 3828 pediatric mental health emergency visits (28,861 clinical notes) to a large health system with 2 EDs (June 2022-October 2024). The reviewers used the Columbia Classification Algorithm for Suicide Assessment to label the presence of SITB at the visit. Random forest classifiers were developed using 3 data modalities: (1) structured data (low-dimensional [International Classification of Diseases codes and chief concerns], medium-dimensional [adding Columbia Suicide Severity Rating Scale screening or mental health diagnoses], and high-dimensional [all structured data or augmented case surveillance, aCS]); (2) text data (general-purpose natural language processing, medical text-specific trained natural language processing, and Large Language Model Meta AI−derived scores), and (3) hybrid data (combining aCS with each text approach). Model performance was evaluated using area under the receiver operating characteristic curve (AUROC).

Results: Of the 3828 visits, 1760 (n=1760, 46.0%) were SITB-related. Detection performance improved with dimensionality: low-dimensional (AUROC=0.865), medium-dimensional (AUROC=0.934 - 0.935), and high-dimensional (AUROC=0.965). Low-dimensional structured (International Classification of Diseases codes and chief concerns) showed high variability in detection, with lower accuracy among preadolescents (AUROC=0.821 vs 0.880 for adolescents); male participants (AUROC=0.817 vs 0.902 for females); and patients with neurodevelopmental (AUROC=0.568 - 0.809), psychotic (AUROC=0.718), and disruptive disorders (AUROC=0.703). Hybrid modality (aCS+Large Language Model Meta AI) achieved optimal performance (AUROC=0.977), with AUROC ≥0.90 for all 20 demographic and 12/15 diagnostic subgroups.

Conclusions: This cross-sectional retrospective study identified that, relative to diagnostic codes and chief concern alone, hybrid structured-text detection methods improved accuracy and mitigated unwanted detection variability. The findings offer a scaffold for future clinical deployment of improved information retrieval of pediatric suicide and self-harm−related emergencies.

(JMIR Ment Health 2026;13:e82371) doi:10.2196/82371

KEYWORDS

children; classification algorithms; electronic medical records; emergency services; mental disorders; natural language processing.; text classification

Introduction

Suicide is the second-leading cause of death among US children over 9 years old [1]. The estimated annual cost of suicide- and self-harm−related emergency department (ED) use is $510 billion, and among young people, nearly 75% of costs are attributable to nonfatal self-harm injuries [2]. Self-injurious thoughts and behaviors (SITB)—encompassing suicidal ideation, suicide attempts, and nonsuicidal self-injury—rank among the strongest predictors of future suicidal behavior [3-5]. The accurate detection of ED visits for SITB underpins interventions to improve quality and reduce preventable ED use [6-8]. Detection enables public health surveillance for geographically or temporally clustered events [7,9-11], informs health system staffing [12], mitigates crowding [13], and supports policy measures such as firearm safety regulations [14,15] and crisis hotlines [16]. Yet, among children, detection remains inconsistent [10,17-19] and leaves many instances of SITB care unidentified [20], particularly among younger children [21].

Several challenges impede the detection of pediatric service use for SITB. When, where, and whether clinicians document suicidality in structured data or clinical text may reflect medical record software functionality [22], stigma [23], racial bias [24], and provider training in pediatrics or mental health (MH) [25,26]. Diagnostic codes and chief concern may inconsistently reflect suicidality in school-age children [21] and children with neurodiverse [27] who often present to emergency services with less lethal means, without immediate disclosure of suicidality, or with externalizing symptoms [28]. The assignment of a diagnostic code often occurs under associated psychiatric diagnoses [29], such as major depression or behavioral disturbance in autism. Diagnostic inaccuracy may further obfuscate these patterns: fewer than 16% of children who attempt suicide are evaluated by a MH specialist in the ED [30].

In this context, methods lag to detect SITB-related service use among children. Most work focuses on adults [31,32] and leverages costly locally trained natural language processing (NLP) of clinical text to detect SITB events in a research context [17]. These NLP methodologies include deep learning [33,34], pretrained models (eg, Word2Vec) [31], and Bidirectional Encoder Representations from Transformers–based transformer models [35] and the examination of keyword representation in clinical notes of individuals with and without self-harm events [31]. While large language models demonstrate promising capabilities to accelerate the efficiency of clinical text analysis, fewer than 5% of medical NLP applications evaluate large language models against nonsynthetic clinical notes using large human−labeled datasets to assess sensitivity, hallucinations, and algorithmic bias [36]. Structured data—such as standardized pediatric MH codes [37] and triage screening [38]—offer more readily implementable detection strategies for operational use [20]. Although the NLP of clinical notes yields fair performance in adolescents [27,39-41], current literature lacks systematic head-to-head comparisons of SITB detection accuracy across electronic health record (EHR) data modalities (text alone, structured alone, hybrid combined). Further, despite calls for algorithmic fairness assessment in suicide prevention [42], phenotyping strategies have seldom evaluated unwanted detection accuracy variation across pediatric demographic and diagnostic subgroups [17]. Combined with typically small human−labeled validation samples (≤1000 youth) [17,32], performance variation in detection strategies across demographic subgroups remains largely unknown.

To address these gaps, this study presents the first large-scale comparative evaluation of automated detection approaches for SITB-related emergency service use among children and adolescents. The primary objectives were to (1) compare detection accuracy across 3 EHR data modalities—structured data alone, clinical text alone, and hybrid combinations—for identifying SITB-related pediatric ED visits; and (2) for each data modality, measure variability in detection performance by youth demographics and underlying MH diagnosis. The findings provide strategies for SITB detection in pediatric emergency settings, with particular emphasis on measuring accuracy for population subgroups historically characterized by suboptimal suicide prevention care.

Methods

Study Design and Population

This retrospective cross-sectional study utilized EHR data from 4 hospitals within a large academic health care system in Southern California serving 5.1 million members, including approximately 400,000 youth. We included all youth aged 6 - 17 years with at least 1 MH-related ED visit between October 2017 and October 2019; this period was selected to capture data following the initial implementation of Columbia Suicide Severity Rating Scale (c-SSRS) screening and *International Classification of Diseases Version 10, Clinical Modification* (*ICD-10-CM*) while excluding pandemic disruptions. MH-ED visits were defined as those associated with (1) a pediatric MH disorder as specified per the Child and Adolescent Mental Health Disorders Classification System (CAMHD-CS), a comprehensive taxonomy organizing pediatric MH-related *ICD-10-CM* codes into diagnostic categories based on *DSM-5* criteria [37]; (2) an MH-related chief concern; (3) involuntary psychiatric detainment; or (4) a positive response to ED nursing triage screening for psychiatric complaints. The flowchart for study inclusion is presented in Multimedia Appendix 1.

To ensure the dataset included unique individuals, we analyzed each child's most recent visit. The multiexpert annotation of all eligible encounters (N=3828 visits) occurred in June 2022-October 2024, with analyses conducted in November 2024-February 2025. We compared 3 data modalities to identify optimal approaches for SITB detection: (1) structured data from discrete EHR fields, (2) text data from clinical narratives, and (3) hybrid combinations integrating all available structured data with NLP of clinical notes. Performance was evaluated against expert classifications using area under the receiver operating characteristic (AUROC) curve metrics for overall cohort and subgroup analyses.

Ethical Considerations

Data, including clinical note text, were deidentified. Analyses were conducted in secure computing environments. The study followed Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) [43] statement guidelines, as well as the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines, TRIPOD-AI [44], and TRIPOD-LLM [45] (Checklist 1). The University of California Los Angeles Institutional Review Board approved this study with the informed consent waiver due to the retrospective study nature and minimal risk (IRB #20-001512).

Demographic and Clinical Variables

Participants were classified as children (6 - 12 y) or adolescents (13 - 17 y), with race, ethnicity, and legal sex from patient- or family-reported EHR. Racial and ethnic categories aligned with federal standards [46]: American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Pacific Islander, White, plus other or missing or unknown. We incorporated 2 area-based socioeconomic measures linked by census tract: the social vulnerability index [47], a Centers for Disease Control and Prevention measure ranking communities' resilience to external stresses on human health (ranging 0 - 1, higher indicating greater vulnerability) derived from the 2018 5-year American Community Survey, and the area deprivation index [48], a composite measure of neighborhood socioeconomic disadvantage based on income, education, employment, and housing quality (national percentile 1 - 100, higher indicating greater deprivation) derived from the 2019 Block Group ADI files v. 3.0. Further details are provided in Multimedia Appendix 2.

We extracted all available EHR data from index ED visits, restricting to the time window between arrival and discharge, transfer, or inpatient hospitalization. We included all verbatim clinical notes (excluding surgical procedures and medical student notes) across disciplines, including notes authored by ED physicians, psychiatrists, psychologists, other medical consultants, social workers, and nurses. We categorized suicide- and self-harm−related diagnoses by the Centers for Disease Control and Prevention surveillance definition of nonfatal suicide attempt and intentional self-harm using *ICD-10-CM* [29]. Youth MH diagnoses were classified via CAMHD-CS, with psychiatric comorbidity defined as ≥2 categories. Chief concerns were categorized as MH or non-MH and SITB or non-SITB related (Multimedia Appendix 3). We included c-SSRS screening scores, ED-administered psychotropic medications, homicidal ideation screening, overdose-related laboratory tests, urine drug screen results, and discharge disposition. Missing data occurred in ≤6% cases for most variables, except for insurance status (~30% missing) and c-SSRS scores (Multimedia Appendix 4). c-SSRS is asked with gatekeeping question structure where subsequent items are only administered if initial screening questions indicate risk. Thus, missingness was recoded as a separate binary variable for each c-SSRS item to preserve this clinical decision pattern. To estimate prior care use, we included number of ED visits and psychiatric and general medical hospitalizations in the past 30, 90, and 365 days.

Ground-Truth Labeling

A total of 2 trained staff research associate annotators reviewed structured data and verbatim notes from each visit. Annotators labeled visits for SITB presence or type using a modified Columbia Classification Algorithm for Suicide Assessment [49] (Multimedia Appendix 5). Interannotator agreement was assessed via Cohen kappa. When annotators disagreed, 2 board-certified child psychiatrists reviewed records independently. Consensus discussion resolved clinician disagreements. All encounters (N=3828) received binary SITB−related and categorical SITB-type classifications. For a random 724-encounter subsample, annotators assigned phrase-level labels indicating SITB-related (any), the SITB type (ideation, attempt, preparatory act, or nonsuicidal self-injury), and if the phrase referred to the patient (vs other), present (vs past), and was affirmed (vs negated).

Text Processing Methods

We developed 3 distinct approaches to assign clinical text scores for SITB detection. We provide complete technical specifications, Community Advisory Board consultation details, and prompt engineering protocols in Multimedia Appendices 6 and 7.

The first approach (general-purpose natural language processing [NLP-general]) adapted a semisupervised methodology from common semisupervised approach (PheCAP) [50] through the following sequential steps. All sentences from the 724 held-out encounters were segmented using spaCy, then embedded using the Universal Sentence Encoder CMLM-en-base and indexed using the Annoy approximate nearest neighbor algorithm [51] with angular distance metrics. Then, for each sentence from the remaining 3104 encounters, the K=5 nearest neighbor sentences were retrieved from the labeled training set, and a sentence-level score was computed as the mean from these neighbors. We determined encounter-level scores by averaging the sentence-level scores (k-normalized votes per sentence) across all sentences within the encounter [51].

The second approach (medical text-specific trained natural language processing) employed identical methodology but substituted MedEmbed-small-v0.146 for sentence vectorization to leverage domain-specific medical embeddings.

The third approach (Large Language Model Meta AI [LLaMA]) utilized large language model processing through a multistage implementation. We leveraged the 724 held-out encounters to iteratively develop and improve upon a condition-specific prompt. The prompt includes instructions to output a Likert-type score ranging from −3 (definitely does not contain SITB) to +3 (definitely contains SITB) along with explanatory text as JSON objects. We tested iterations of this prompt using LLaMA-3.2-1B (selected for computational efficiency with 10× faster processing speed) by comparing the Likert scores against note-level labels from human reviewers. Once preliminary accuracy was established, we presented the prompt to the study Community Advisory Board that suggested additional revisions. Once the prompt was finalized, we conducted final scoring on the remaining 3104 encounters' notes using LLaMA-3.3-70B. We then determined encounter-level scores by selecting the maximum score across all clinical notes within each encounter.

Classification Models

Feature Set Definitions

We define 3 data modalities based on the fundamental data type: (1) structured modality used discrete EHR fields, (2) text modality used clinical narratives processed through NLP, and (3) hybrid modality combined both data types. We compared a total of 10 feature sets against multiexpert chart annotation—4 structured modality, 3 text modality, and 3 hybrid modality.

We categorized structured feature sets by dimensionality based on the number of input features: low (<10 features), medium (10 - 50 features), and high (>50 features). The 4 structured data feature sets were as follows: *Low*: (1) SITB-related *ICD-10-CM* codes and chief concerns (International Classification of Diseases codes and chief concerns [ICD/CC]); *Medium*: either (2) low plus c-SSRS scores from ED nursing evaluation (c-SSRS+ICD/CC) or (3) low plus MH diagnoses from primary treating ED physician evaluation (MH dx+ICD/CC); and *High*: (4) augmented case surveillance (aCS), which includes all available structured clinical data from the EHR. We categorized feature sets by dimensionality to understand the trade-off between model complexity and performance, where low-dimensional models are easier to implement but may miss important signals.

The 3 text feature sets were (5) NLP-general, (6) NLP-med, and (7) an open-source large language model (LLaMA). We selected these text approaches to evaluate detection gains while accounting for key trade-offs—dependency on sentence-labeled data (yes: NLP-general and NLP-med, no: LLaMA), computational resource requirements (higher graphics processing unit requirements: NLP-med, LLaMA; higher central processing unit requirements: NLP-general), and medical-specific versus light-weight embeddings (NLP-med vs NLP-general).

The 3 hybrid structured-text feature sets (8-10) combined aCS with each text approach.

Model Development and Validation

Encounters allocated to develop text processing methodology (n=724) were excluded. To assign the probabilities of SITB presence to the remaining encounters (n=3104), we developed 10 random forest classifiers [52], using 10-fold cross-validation with nested hyperparameter optimization [53]. A probability threshold of 0.5 was applied to convert random forest predictions into binary encounter-level classifications. For each outer fold, training data was split 50/50 for inner cross-validation. Hyperparameters were selected from the grid based on the highest classification accuracy in the inner CV. Each outer fold could select different optimal hyperparameters independently. The AUROC was calculated separately for each outer fold using the selected hyperparameters. The reported AUROC values represent the mean across all 10 outer folds with 95% CIs. The mean receiver operating characteristic (ROC) curve was created by interpolating individual fold ROC curves onto a common false positive rate grid and averaging the true positive rates. We selected this approach to maintain the integrity of the validation process and prevent data leakage by ensuring that hyperparameter tuning occurs only on training folds, with performance evaluation conducted on completely unseen validation data within each fold. Each encounter classifier's individual features are specified in Multimedia Appendix 8.

Statistical Analysis

Overall Classification Performance

We evaluated performance using AUROC, accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. Shapley Additive Explanation values quantified feature importance, while permutation importance provided complementary ranking. Cross-validation variability was used to construct asymptotically exact CIs for test error [54]. Classifier performance was compared using DeLong tests [55].

Subgroup Performance

We assessed subgroup variation [56] by stratifying performance across demographic (age group, sex, race or ethnicity) and MH diagnosis (CAMHD-CS groups) subgroups. Each patient was assigned to 1 demographic subgroup but could belong to multiple diagnostic subgroups. For each subgroup, we calculated performance metrics with 95% CIs and generated ROC curves.

Analyses used Python 3.13.0 with scikit-learn, pytorch 1.9.0, and spacy 3.2.0. LLaMA inferencing used Hugging Face Transformers v4.49.0. The code is available upon request.

Results

Sample Characteristics

Our study sample included 3828 pediatric ED visits by unique youth ages 6 - 17 and comprised 28,861 notes with 619,827 sentences. The sample included 1963 (51.3%) female and 1865 (48.7%) male youth, with the racial and ethnic composition of White non-Hispanic (n=1894, 49.5%), Hispanic or Latino (n=1017, 26.6%), Black (n=363, 9.5%), and Asian (n=178, 4.6%; Table 1). Adolescents (ages 13 - 17 y) constituted most of the sample (n=2819, 73.6%), while children (ages 6 - 12 y) represented 26.4% (n=1009). The median age was 15 (IQR 12 - 16) years. Common psychiatric diagnoses included depressive disorders (n=1387, 36.2%), anxiety disorders (n=1161, 30.3%), suicide or self-injury coded diagnoses (n=1282, 33.5%), and attention-deficit or hyperactivity disorder (ADHD) (n=840, 21.9%). Suicide-related concerns comprised 18.5% (n=708) of the chief concerns.

Table . Sample characteristics[a].

Sample characteristic	Value
Gold-standard, n (%)	3828 (100)
Any SITB[b]	1760 (46.0)
Suicide attempt	301 (7.9)
Preparatory acts	261 (6.8)
Suicidal ideation	1014 (26.5)
NSSI[c]	762 (19.9)
Other reason for visit	2036 (53.2)
Not enough information	33 (0.9)
Sex, n (%)	
Female	1963 (51.3)
Male	1865 (48.7)
Race and ethnicity, n (%)	
Not Hispanic or Latino	2774 (72.5)
American Indian or Alaska Native	10 (0.3)
Asian	178 (4.6)
Black or African American	363 (9.5)
Multiple races	88 (2.3)
Native Hawaiian or Other Pacific Islander	5 (0.1)
White	1894 (49.5)
Other race	235 (6.1)
Hispanic or Latino	1017 (26.6)
Unknown race or ethnicity	37 (1.0)
Site, n (%)	
Academic medical center	2858 (74.7)
Community hospital	970 (25.3)
Disposition, n (%)	
Discharged without hospitalization	2277 (59.5)
Hospitalized	1452 (37.9)
General medical hospitalization	390 (10.2)
Psychiatric hospitalization	1062 (27.7)
Other disposition	99 (2.6)
Chief concern, n (%)	
Psychiatric (including suicide-related)	2108 (55.1)
Suicide-related	708 (18.5)
ED[d] Diagnostic code category (CAMHD-CS[e]), n (%)	
ADHD[f]	840 (21.9)
Anxiety disorders	1161 (30.3)
Autism spectrum disorder	468 (12.2)
Bipolar and related disorders	176 (4.6)
Depressive disorders	1387 (36.2)
Developmental disorder	81 (2.1)

Sample characteristic	Value
Disruptive, impulse control, and conduct disorders	269 (7.0)
Feeding and eating disorders	106 (2.8)
Intellectual disability	66 (1.7)
Mental health symptom	535 (14.0)
Miscellaneous	202 (5.3)
Neurocognitive disorders	66 (1.7)
Obsessive-compulsive and related disorders	172 (4.5)
Schizophrenia and other psychotic disorders	145 (3.8)
Substance-related and addictive disorders	475 (12.4)
Suicide or self-injury	1282 (33.5)
Trauma and stressor-related disorders	246 (6.4)
≥2 CAMHD-CS[e] diagnoses	2345 (61.3)
Age (y), median (IQR)	15 (12 - 16)
Social vulnerability index, total, median (IQR)	0.38 (0.19 - 0.65)
Area deprivation index, median (IQR)	
State ranking	2 (1-5)
National ranking	5 (2-12)

[a]Percentages do not sum to 100% as children may present with more than one chief concern or mental health diagnosis.
[b]SITB: self-injurious thoughts and behaviors.
[c]NSSI: nonsuicidal self-injury.
[d]ED: emergency department.
[e]CAMHD-CS: Child and Adolescent Mental Health Disorders Classification System.
[f]ADHD: attention-deficit or hyperactivity disorder.

Ground-Truth Agreement

The raters agreed on SITB classification (3695/3828 [96.5% agreement]; Cohen κ=0.93). Nearly half (n=1760, 46.0%) of the encounters involved SITB, with similar prevalence in children (n=455, 45.1%) and adolescents (n=1305, 46.3%).

Performance Metrics

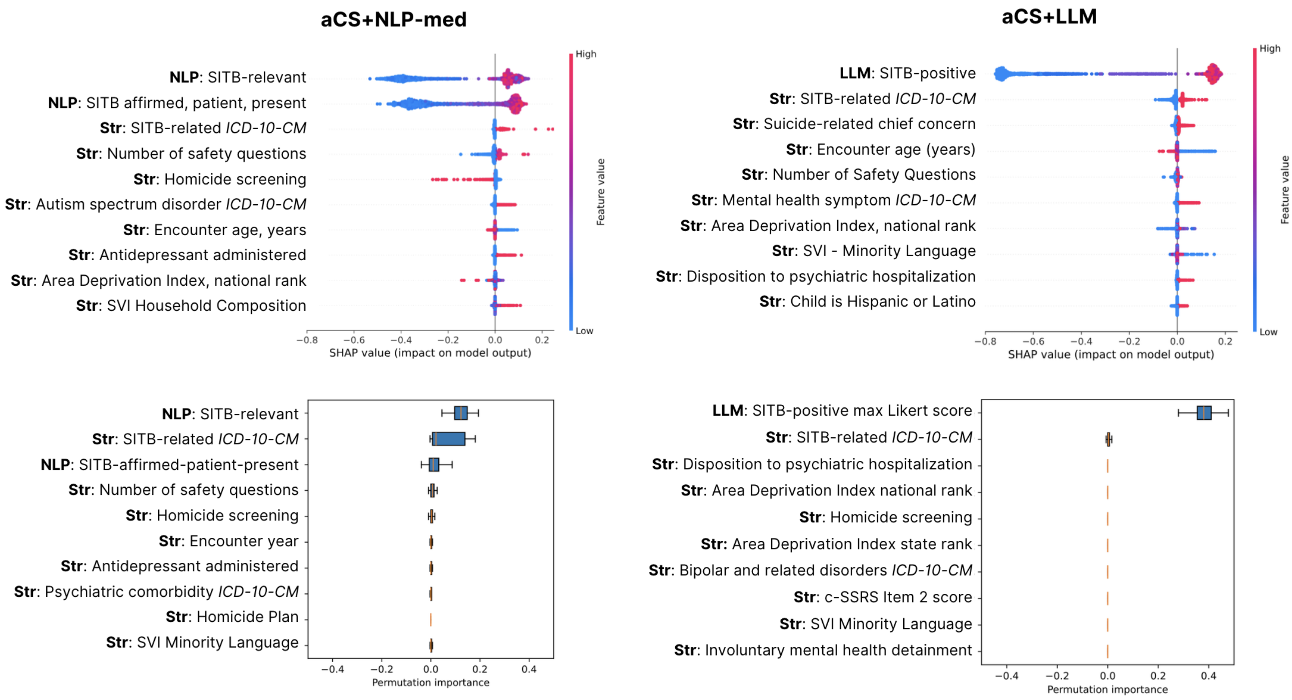
Overview

The detection of SITB varied across EHR data representations (Figure 1). Complete fit metrics, failure mode characterization including the number of false positives and false negatives, exact DeLong test *P* values, and ROC curves are provided in Multimedia Appendices 9 and 10. AUROC stability across folds is visualized in Multimedia Appendix 11. The examination of Shapley Additive Explanations and permutation importances for the best-performing representations revealed that text-derived features provided the strongest contribution to classifier accuracy (Figure 2; Multimedia Appendices 12 and 13).

Figure 1. Comparison of detection classifiers for self-injurious thoughts and behaviors: pairwise analysis of area under the receiver operating curve (AUROC) by feature type. This heatmap illustrates the differences in AUROC values among 10 feature sets used for encounter classification. The matrix compares each pair of feature sets, with subtrahend feature sets listed in columns and minuend feature sets listed in rows. Cells shaded blue indicate that the row feature set outperformed the column feature set, while cells shaded yellow indicate inferior performance. The feature sets are categorized into 3 groups: structured (including International Classification of Diseases codes and chief concerns [ICD/CC], c-SSRS+ICD/CC, MH dx+ICD/CC, and augmented case surveillance [aCS]), text (including general-purpose natural language processing [NLP-general], medical text-specific trained natural language processing [NLP-med], and Large Language Model Meta AI [LLaMA]), and hybrid (including aCS+NLP, aCS+NLP-med, and aCS+LLaMA). Asterisks denote statistical significance (**P*<.05, ***P*<.01, ****P*<.001). The largest improvement in AUROC (0.111) occurred when high-dimensional structured data (aCS) were combined with open-source large language model scores (LLaMA), compared to the baseline ICD/CC feature set. c-SSRS: Columbia Suicide Severity Rating Scale.

		Subtrahend									
			Structured data			**Note text**			**Hybrid**		
	Feature set	ICD/CC	c-SSRS+ ICD/CC	MH dx+ ICD/CC	aCS	NLP-gen	NLP-med	LLaMA	aCS+ NLP-gen	aCS+ NLP-med	*Difference Color Map*
Minuend **Structured data**	ICD/CC										0.100
	c-SSRS+ICD/CC	0.069***									0.075
	MH dx+ICD/CC	0.068***	-0.001								0.050
	aCS	0.100***	0.030***	0.032***							0.025
Text data	NLP-gen	0.091***	0.021***	0.022***	-0.009**						0.000
	NLP-med	0.105***	0.035***	0.037***	0.005	0.014***					-0.025
	LLaMA	0.097***	0.027***	0.029***	-0.003	0.006	-0.008*				-0.050
Hybrid	aCS+NLP-gen	0.107***	0.038***	0.039***	0.007***	0.017***	0.002	0.010**			-0.075
	aCS+NLP-med	0.106***	0.037***	0.038***	0.007**	0.016***	0.002	0.010**	-0.001		-0.100
	aCS+LLaMA	0.111***	0.042***	0.043***	0.012***	0.021***	0.007*	0.015***	0.004*	0.005*	

Figure 2. Comparison of feature and permutation importance between aCS+NLP-med and aCS+LLaMA for encounter classification of self-injurious thoughts and behaviors. This figure presents importance analyses for the 2 top-performing classification approaches, aCS+NLP-med (left panels) and aCS+LLaMA (right panels), which achieved the highest area under the receiver operating characteristic curve. The upper panels display Shapley Additive Explanation (SHAP) values, where negative values indicate the decreased detection of self-injurious thoughts and behaviors (SITB), and positive values indicate increased detection. The lower panels show permutation importance scores, which quantify the contribution of each feature to the classifier's performance. In both classifiers, text features dominated the feature importance rankings, outperforming structured data features. Notable exceptions among structured data features included Columbia Suicide Severity Rating Scale (c-SSRS) items, homicide screening, area deprivation indices, encounter age, and psychiatric hospitalization disposition. For reference, the notation used in this figure is as follows: aCS represents all available structured data; NLP-med refers to note scores derived using MedEmbed-small-v0.1 embeddings with nearest-neighbor approximation; LLM refers to note scores generated by the open-source language model llama-3.3-70B. The "+" symbol indicates combinations of aCS with the corresponding text-based feature set (NLP-med or Large Language Model Meta AI [LLaMA]). aCS: augmented case surveillance; *ICD-10-CM*: *International Classification of Diseases Clinical Modification Version 10*; NLP-med: medical text-specific trained natural language processing; Str: structured data; SVI: social vulnerability index.



Structured Data Classification

Low-dimensional structured (ICD/CC) yielded the lowest accuracy detection (AUROC 0.865, 95% CI 0.852 - 0.879). Both medium-dimensional structured (c-SSRS+ICD/CC; MH dx+ICD/CC) feature sets outperformed ICD/CC (both *P*<.001).

The high-dimensional structured feature set (aCS) (AUROC 0.965, 95% CI 0.958 - 0.972) further outperformed c-SSRS+ICD/CC (AUROC 0.935, 95% CI 0.925 - 0.944) and MH dx+ICD/CC (AUROC 0.934, 95% CI 0.924 - 0.943; *P*<.001).

Text-Based Classification

Among text modalities, NLP-med (AUROC 0.970, 95% CI 0.964 - 0.977) marginally outperformed NLP-general (AUROC 0.956, 95% CI 0.948 - 0.964; *P*<.001) and LLaMA (AUROC 0.962, 95% CI 0.955 - 0.969; *P*=.03). The text-only modalities all surpassed ICD/CC as well as c-SSRS+ICD/CC and MH dx+ICD/CC (all *P*<.001), but NLP-general was slightly inferior to aCS (*P*=.005). NLP-med and LLaMA did not significantly exceed aCS.

Hybrid Classification

Combining text with aCS exceeded aCS alone (*P*<.001). Combining text with aCS also exceeded NLP-general alone (*P*<.001) and LLaMA alone (*P*<.01). However, adding aCS to NLP-med did not improve detection compared with NLP-med alone (*P*=.633). The hybrid representation combining aCS with LLaMA classification (aCS+LLaMA) achieved the highest overall AUROC (0.977, 95% CI 0.971 - 0.982), narrowly exceeding aCS with NLP-med (AUROC 0.970, 95% CI 0.964 - 0.977; *P*=.04).

Subgroup Performance

Demographic Subgroups

Detection varied considerably across age, sex, and race or ethnicity subgroups (Figure 3; Multimedia Appendix 14). Low-dimensional structured data (ICD/CC) achieved AUROC values ≥0.950 for only 2/20 demographic groups. ICD/CC performed less well for children (AUROC 0.821, 95% CI 0.791 - 0.851) compared to adolescents (AUROC 0.880, 95% CI 0.865 - 0.895) and for male (AUROC 0.817, 95% CI 0.794 - 0.840) compared to female (AUROC 0.902; 95% CI 0.886 - 0.918) youth, with nonoverlapping CIs. Detection was similar between female children and female adolescents (AUROC 0.903, 95% CI 0.867 - 0.939 vs AUROC 0.902, 95% CI 0.884 - 0.920) but differed between male children and male adolescents (AUROC 0.753, 95% CI 0.708 - 0.798 vs AUROC 0.847, 95% CI 0.821 - 0.873). Using ICD/CC alone, detection was lower among Hispanic male children (AUROC 0.684, 95% CI 0.579 - 0.789) and Black male children (AUROC 0.754, 95% CI 0.630 - 0.877), with a similar trend among Asian male children. Multimedia Appendices 16-19 present ROC curves stratified by feature set and demographic groups.

Figure 3. Stratified performance of detection classifiers by demographics. This figure presents the area under the receiver operating characteristic curve (AUROC) values for various encounter classification feature sets, stratified by age (6-12 and 13-17 y), sex (male and female), and race or ethnicity (Asian, Black, Hispanic or Latino, White, and Other), with the number of cases with self-injurious thoughts and behaviors (SITB Pos) shown for each subgroup. The feature sets are categorized into 3 groups: structured (including International Classification of Diseases codes and chief concerns [ICD/CC], c-SSRS+ICD/CC, MH dx+ICD/CC, and augmented case surveillance [aCS]), text (including general-purpose natural language processing [NLP-gen], medical text-specific trained natural language processing [NLP-med], and Large Language Model Meta AI [LLaMA]), and hybrid (including aCS+NLP, aCS+NLP-med, and aCS+LLaMA). The results show that baseline classifiers using only ICD codes or chief concerns had lower performance (AUROC range: 0.681-0.966), whereas more comprehensive classifiers, particularly those combining structured data (aCS) with natural language processing (NLP or LLaMA), achieved higher performance across all demographic subgroups (AUROC range: 0.900-1.000), as indicated by the color gradient from teal (higher performance) to red (lower performance), with the "Other" race or ethnicity category including individuals who identify as multiple races, Native Hawaiian or Pacific Islander, Native American or Alaska Native, or have an unknown race or ethnicity.

			Structured data				Note text			Hybrid		
Demographic Group	*N*	*n SITB, %*	ICD/CC	c-SSRS +ICD/CC	MH dx +ICD/CC	aCS	NLP-gen	NLP-med	LLaMA	aCS+ NLP-gen	aCS+ NLP-med	aCS+ LLaMA
Male, Asian, 6-12 years	18	6 (33.3)	0.681	0.84	0.903	0.833	0.875	0.917	0.986	0.875	0.889	0.972
Male, Latino, 6-12 years	144	37 (25.7)	0.684	0.832	0.871	0.956	0.972	0.962	0.956	0.975	0.976	0.965
Male, Black, 6-12 years	61	29 (47.5)	0.754	0.848	0.921	0.921	0.934	0.949	0.932	0.918	0.927	0.958
Male, White, 6-12 years	233	107 (45.9)	0.775	0.843	0.916	0.912	0.911	0.941	0.952	0.926	0.941	0.963
Female, Other Race/Ethnicity, 6-12 years	30	13 (43.3)	0.785	0.928	0.91	0.959	0.986	1	1	0.968	0.995	1
Male, Latino, 13-17 years	262	85 (32.4)	0.802	0.918	0.908	0.953	0.939	0.95	0.94	0.963	0.956	0.956
Male, Other Race/Ethnicity, 13-17 years	74	22 (29.7)	0.838	0.939	0.873	0.959	0.965	0.986	0.968	0.969	0.978	0.976
Male, Black, 13-17 years	80	22 (27.5)	0.841	0.943	0.969	0.971	0.931	0.972	0.934	0.969	0.97	0.985
Male, Other Race/Ethnicity, 6-12 years	48	11 (22.9)	0.844	0.915	0.971	0.958	0.968	0.966	0.983	0.988	0.962	0.988
Female, Black, 6-12 years	34	13 (38.2)	0.846	0.949	0.813	0.967	0.96	0.993	0.978	0.974	0.989	0.963
Female, Black, 13-17 years	113	51 (45.1)	0.848	0.917	0.919	0.958	0.928	0.958	0.952	0.964	0.962	0.973
Male, White, 13-17 years	560	244 (43.6)	0.849	0.941	0.919	0.957	0.941	0.963	0.957	0.969	0.964	0.971
Female, Latino, 13-17 years	311	154 (49.5)	0.887	0.949	0.933	0.973	0.973	0.985	0.964	0.978	0.975	0.978
Female, White, 13-17 years	618	325 (52.6)	0.912	0.959	0.964	0.977	0.96	0.977	0.969	0.978	0.98	0.981
Female, White, 6-12 years	137	71 (51.8)	0.915	0.95	0.933	0.966	0.974	0.971	0.973	0.972	0.961	0.98
Female, Asian, 6-12 years	16	10 (62.5)	0.925	0.967	1	0.967	0.933	0.892	0.883	0.917	0.95	0.9
Female, Latino, 6-12 years	110	48 (43.6)	0.929	0.98	0.967	0.997	0.998	1	0.998	0.999	0.998	0.997
Female, Other Race/Ethnicity, 13-17 years	105	49 (46.7)	0.932	0.93	0.941	0.984	0.971	0.976	0.968	0.989	0.981	0.992
Female, Asian, 13-17 years	63	32 (50.8)	0.946	0.979	0.981	0.993	0.99	0.976	0.984	0.997	0.996	0.996
Male, Asian, 13-17 years	40	21 (52.5)	0.966	0.995	0.951	0.997	0.98	0.98	0.961	1	0.997	1

Hybrid feature sets achieved the greatest subgroup consistency, with AUROC values ≥0.950 for aCS+NLP-general (15/20 demographic groups), aCS+NLP-med (17/20 demographic groups), and aCS+LLaMA (19/20 demographic groups). aCS+LLaMA yielded the most consistent performance, reducing the AUROC gap between the highest and lowest performing groups from 0.285 (ICD/CC) to 0.100 (aCS+LLaMA). Notably, aCS+LLaMA achieved strong detection performance for groups with the lowest detection performance using ICD/CC alone, including Hispanic male children and Black male children, with AUROC improvements of 0.281 and 0.205, respectively.

Diagnostic Subgroups

Detection further varied by MH diagnostic categories (Figure 4; Multimedia Appendix 19). ICD/CC achieved AUROC values ≥0.90 for 0/15 diagnostic groups. ICD/CC achieved lower SITB detection performance among youth with neurodevelopmental (eg, intellectual disability: AUROC 0.568, 95% CI 0.410 - 0.726; autism spectrum disorder: AUROC 0.736, 95% CI 0.686 - 0.787; ADHD: AUROC 0.809, 95% CI 0.777 - 0.841), externalizing (disruptive or impulse control disorders: AUROC 0.703, 95% CI 0.635 - 0.772), and psychotic (AUROC 0.718, 95% CI 0.626 - 0.811) disorders. In contrast, ICD/CC achieved higher SITB detection performance among youth with internalizing (eg, depressive disorders: AUROC 0.896, 95% CI 0.878 - 0.915; anxiety disorders: AUROC 0.878, 95% CI 0.856 - 0.899; trauma- or stressor-related disorders: AUROC 0.842, 95% CI 0.789 - 0.895) and substance-related disorders (AUROC 0.878, 95% CI 0.840 - 0.917).

Figure 4. Stratified performance of detection classifiers by mental health diagnosis. This figure presents the area under the receiver operating characteristic curve (AUROC) values for various encounter classification feature sets, stratified by the 15 most prevalent diagnostic categories of the Child and Adolescent Mental Health Disorders Classification System (CAMHD-CS), with the number of self-injurious thoughts and behaviors (SITB)−positive cases shown for each subgroup. For performance by all 23 categories, see Multimedia Appendix 13. The feature sets are categorized into 3 groups: structured (including International Classification of Diseases codes and chief concerns [ICD/CC], c-SSRS+ICD/CC, MH dx+ICD/CC, and augmented case surveillance [aCS]), text (including general-purpose natural language processing [NLP-gen], medical text-specific trained natural language processing [NLP-med], and Large Language Model Meta AI [LLaMA]), and hybrid (including aCS+NLP, aCS+NLP-med, and aCS+LLaMA). The results show that classifiers performed best in identifying SITB risk for substance-related and addictive disorders, anxiety disorders, and developmental delay disorders across most feature sets. Notably, classifiers that integrated structured data (aCS) with natural language processing (NLP) or large language model (LLaMA) approaches generally achieved higher performance compared to individual feature sets alone, as indicated by the color gradient from teal (higher performance) to red (lower performance). ADHD: attention-deficit or hyperactivity disorder.

			Structured data				Note text			Hybrid		
Diagnosis	*N*	*n SITB, %*	ICD/CC	c-SSRS +ICD/CC	MH dx +ICD/CC	aCS	NLP-gen	NLP-med	LLaMA	aCS+ NLP-gen	aCS+ NLP-med	aCS+ LLaMA
Intellectual disability	55	21 (38.2)	0.568	0.756	0.768	0.902	0.898	0.902	0.945	0.899	0.913	0.933
Disruptive, impulse control and conduct disorders	436	295 (67.7)	0.703	0.728	0.618	0.743	0.730	0.839	0.859	0.752	0.816	0.858
Schizophrenia spectrum and other psychotic disorders	158	87 (55.1)	0.718	0.777	0.648	0.759	0.750	0.823	0.829	0.801	0.809	0.867
Developmental delay or neurodevelopmental disorder	36	17 (47.2)	0.735	0.912	0.902	0.968	0.952	0.986	0.927	0.966	0.950	0.973
Autism spectrum disorder	54	10 (18.5)	0.736	0.851	0.816	0.902	0.899	0.927	0.924	0.915	0.912	0.936
Mental health symptom	129	89 (69.0)	0.745	0.822	0.661	0.841	0.791	0.846	0.856	0.862	0.875	0.880
Miscellaneous	23	18 (78.3)	0.763	0.864	0.853	0.909	0.913	0.938	0.918	0.937	0.940	0.939
Feeding and eating disorders	116	74 (63.8)	0.774	0.849	0.858	0.909	0.885	0.910	0.898	0.921	0.921	0.952
Bipolar and related disorders	29	25 (86.2)	0.804	0.819	0.792	0.828	0.791	0.911	0.867	0.844	0.876	0.891
ADHD	23	5 (21.7)	0.809	0.895	0.897	0.934	0.932	0.949	0.941	0.950	0.950	0.956
Trauma and stressor-related disorders	11	5 (45.5)	0.842	0.914	0.840	0.934	0.896	0.944	0.917	0.939	0.948	0.936
Obsessive-compulsive disorders	389	152 (39.1)	0.864	0.919	0.839	0.958	0.924	0.947	0.949	0.948	0.942	0.967
Comorbidity (≥ 2 CAMHD-CS diagnostic groups)	202	129 (63.9)	0.877	0.915	0.869	0.931	0.919	0.946	0.931	0.946	0.948	0.957
Anxiety disorders	1866	1286 (68.9)	0.878	0.929	0.903	0.950	0.953	0.963	0.953	0.964	0.962	0.970
Substance related and addictive disorders	700	357 (51.0)	0.878	0.937	0.921	0.971	0.972	0.975	0.958	0.978	0.976	0.974
Depressive disorders	933	574 (61.5)	0.896	0.931	0.879	0.942	0.934	0.957	0.942	0.960	0.964	0.963

Hybrid feature sets achieved the greatest subgroup consistency among diagnostic groups, with AUROC exceeding 0.950 for aCS+NLP-general (5/15 diagnostic groups), aCS+NLP-med (5/15 diagnostic groups), and aCS+LLaMA (8/15 diagnostic groups). aCS+LLaMA yielded the most consistent performance, reducing the AUROC gap between the highest and lowest performing groups from 0.328 (ICD/CC classifier) to 0.115. Notably, aCS+ LLaMA achieved strong detection performance for the groups with lower detection performance using ICD/CC alone, including neurodevelopmental problems (intellectual disability: AUROC 0.933, 95% CI 0.854 - 1.000; developmental delay: AUROC 0.973, 95% CI 0.910 - 1.000; autism spectrum disorder: AUROC 0.936, 95% CI 0.909 - 0.963; ADHD: AUROC 0.956, 95% CI 0.941 - 0.972), externalizing problems (disruptive or impulse control disorders: AUROC 0.858, 95% CI 0.809 - 0.908), and psychotic disorders (AUROC 0.867, 95% CI 0.803 - 0.931). However, the detection of SITB among children with externalizing and psychotic disorders remained lower for internalizing disorders (eg, depression: AUROC 0.963, 95% CI 0.953 - 0.973).

Discussion

Principal Findings

In this cross-sectional study, integrating comprehensive structured data with clinical notes substantially improved the detection of pediatric ED service use for suicide and self-harm. Hybrid modality classifiers combining high-dimensional structured data with an open-source language model scores achieved the highest performance across nearly all subgroups—advancing detection accuracy beyond prior efforts relying on *ICD-10-CM* codes or clinical text alone [17,35,57,58]. Surprisingly, detection using high-dimensional structured data approximated text-based approaches, providing a resource-efficient alternative to improve detection while simplifying anonymization and computational requirements.

Our findings challenge the widespread reliance on suicide- and self-harm−related *ICD-10-CM* codes and chief concern for identifying SITB emergency service use among children. While epidemiologic studies report female adolescents account for the surge in emergency service use for suicidality [59,60], the misclassification of SITB among male children may distort observed patterns of pediatric ED utilization. This detection gap raises particular concern given the annual 8.2% rise in suicide death rates among preteens [59] and the highest age-standardized suicide death rates among male US youth aged 10 - 24 years across 52 countries [61]. Youth with psychotic disorders or neurodevelopmental disorders also presented detection challenges despite their markedly elevated risk—70-fold elevated risk of suicide attempts [62] and 3-fold elevated risk of suicide death [63,64], respectively. For these populations, clinical text analysis offers advantages, possibly by capturing subtle manifestations of distress such as irritability, perceptual disturbances, and aggression. Future phenotyping studies should implement systematic bias auditing protocols that regularly evaluate detection accuracy across demographic and diagnostic subgroups to identify and remediate performance disparities before clinical deployment.

The detection of pediatric ED service use for suicide and self-harm has key implications for clinical practice, health system operations, and public health surveillance. Better detection underpins the development of clinical decision support tools to guide clinician awareness of suicide risk and promote delivery of evidence-based suicide prevention interventions such as safety planning [65,66] and lethal means safety counseling [67,68]. Youth with serious mental illness and developmental disorders—among the most frequently undetected groups—are also the highest ED utilizers [59] and experience extended boarding times [69]. Investment in sensitive, efficient SITB detection methods is likely to yield significant returns through forecasting resources, alleviating ED crowding, and reducing ED recidivism. Further, narrowing detection gaps could enable more precise monitoring during crisis periods, such as natural disasters and suicide clusters.

Limitations

While this study included 2 EDs in a single health system, generalizability requires external validation in other health systems, particularly in low-resourced community-based settings. Some children with SITB present to health care settings for non-MH reasons. To balance maximizing unique individuals while maintaining feasibility for human annotation, we analyzed only the most recent MH-related visit per child. Model performance was evaluated using retrospective data that may not reflect evolving clinical documentation practices, changes in suicide screening protocols, or shifts in patient presentation patterns over time. While feature dimensions are invariant to note length, it is possible extreme differences in documentation volume could influence the accuracy of note-derived scores; however, our observation of decreased heterogeneity with the use of text suggests that note-derived features are capturing clinical patterns across subgroups despite any unmeasured documentation differences. Calibration, while essential for clinical deployment, was outside our scope of comparing data modalities' relative discriminative power. There are numerous practical challenges involved in deploying NLP methodologies in real-time clinical settings, including the computational cost and necessary implementation infrastructure. Future research should focus on prospective validation in diverse clinical settings, implementation studies examining workflow integration and clinician acceptance, cost-effectiveness analyses, and evaluation of model degradation over time. In the interim, this study offers actionable approaches to strengthening retrospective surveillance of pediatric suicide−related ED use. Real-time EHR integration would require robust model maintenance protocols, comprehensive staff training on result interpretation, patient and family input on automated screening approaches, and ongoing bias monitoring.

Conclusions

This study developed a cross-disciplinary and multimodal machine learning approach for automating the detection of pediatric SITB−related emergency care using integrated EHR data representations. The hybrid modality achieved high accuracy while demonstrating reduced variation across demographic and diagnostic subgroups compared with basic structured data alone. The findings indicate that, alone, *ICD-10-CM* codes and chief concerns yield suboptimal and variable detection accuracy. Study methods provide computationally efficient alternatives to improve detection accuracy beyond traditional approaches. The findings suggest that systematic detection gaps exist and can be efficiently mitigated: focused efforts to augment information retrieval on suicide risk factors at bedside are needed to stymie decision bias and bolster pediatric MH care quality.

Acknowledgments

We would like to thank the members of our Community Advisory Board for their valuable input on prompt refinement. We are also grateful to the UCLA Biomedical Informatics Program for assistance with medical record data extraction and area deprivation index or social vulnerability index linkage. This study utilized LLaMA-3.2-1B and LLaMA-3.3-70B for text analysis as described in the *Methods* section.

Funding

This work was supported by the National Institute of Mental Health (K23-MH130745), the Harvey T. and Maude C. Sorensen Foundation, and National Center for Advancing Translational Sciences of the National Institutes of Health (UL1TR001881) (Edgcomb).

Data Availability

The use of protected health information from verbatim clinical mental health notes entails strict confidentiality and privacy regulations to safeguard the sensitive information of the individuals, in accordance with legal and ethical standards governing health care data. The nature of our dataset containing protected health information prohibits its release to maintain compliance with regulatory requirements and to uphold patient confidentiality.

Authors' Contributions

Conceptualization: BTZ, JBE
Data curation: JBE
Formal analysis: AK, AS, JBE, JJL
Funding acquisition: JBE
Investigation: CGP, ET, JBE, JJL, TT
Methodology: AK, AS, BTZ, JBE
Project administration: CGP, ET
Software: JJL, TT
Supervision: BTZ, JBE
Writing original draft: JBE
Writing review & editing: AK, AS, BTZ, CGP, ET, JBE, JJL, TT

Conflicts of Interest

None declared.

Multimedia Appendix 1
Flow diagram for study inclusion.
[DOCX File, 210 KB - mental_v13i1e82371_app1.docx]

Multimedia Appendix 2
Variable construction.
[DOCX File, 17 KB - mental_v13i1e82371_app2.docx]

Multimedia Appendix 3
Mental health−related chief complaints in structured data fields.
[DOCX File, 14 KB - mental_v13i1e82371_app3.docx]

Multimedia Appendix 4
Methods to handle missingness.
[DOCX File, 13 KB - mental_v13i1e82371_app4.docx]

Multimedia Appendix 5
Chart annotation guide.
[DOCX File, 33 KB - mental_v13i1e82371_app5.docx]

Multimedia Appendix 6
Software implementation.
[DOCX File, 34 KB - mental_v13i1e82371_app6.docx]

Multimedia Appendix 7
Community advisory board.
[DOCX File, 15 KB - mental_v13i1e82371_app7.docx]

Multimedia Appendix 8
Variables comprising feature sets.
[DOCX File, 28 KB - mental_v13i1e82371_app8.docx]

Multimedia Appendix 9
Detection performance by classifier.
[DOCX File, 18 KB - mental_v13i1e82371_app9.docx]

Multimedia Appendix 10
DeLong tests (*P* value) by feature set.
[DOCX File, 17 KB - mental_v13i1e82371_app10.docx]

Multimedia Appendix 11
Area under the receiver operating characteristic (AUROC) curves by feature set.
[DOCX File, 582 KB - mental_v13i1e82371_app11.docx]

Multimedia Appendix 12
Shapley Additive Explanations plots by feature set.
[DOCX File, 530 KB - mental_v13i1e82371_app12.docx]

Multimedia Appendix 13
Permutation feature importance by feature set.
[DOCX File, 366 KB - mental_v13i1e82371_app13.docx]

Multimedia Appendix 14
Detection performance by classifier and age group and by sex and race or ethnicity.
[DOCX File, 196 KB - mental_v13i1e82371_app14.docx]

Multimedia Appendix 15
Area under the receiver operating characteristic curves (AUROC) by feature set by sex and age group (children).
[DOCX File, 377 KB - mental_v13i1e82371_app15.docx]

Multimedia Appendix 16
Area under the receiver operating characteristic (AUROC) curves by feature set by sex and age group (adolescents).
[DOCX File, 377 KB - mental_v13i1e82371_app16.docx]

Multimedia Appendix 17
Area under the receiver operating characteristic (AUROC) curves by feature set by race or ethnicity (children).
[DOCX File, 482 KB - mental_v13i1e82371_app17.docx]

Multimedia Appendix 18
Area under the receiver operating characteristic (AUROC) curves by feature set by race or ethnicity (adolescents).
[DOCX File, 494 KB - mental_v13i1e82371_app18.docx]

Multimedia Appendix 19
Detection performance by classifier and emergency department mental health diagnosis.
[DOCX File, 72 KB - mental_v13i1e82371_app19.docx]

Checklist 1
STROBE and TRIPOD checklists.
[DOCX File, 28 KB - mental_v13i1e82371_app20.docx]

References

1. Centers for disease control and prevention. WISQARS Data Visualization. URL: https://wisqars.cdc.gov/ [accessed 2026-01-29]
2. Peterson C, Haileyesus T, Stone DM. Economic cost of US suicide and nonfatal self-harm. Am J Prev Med 2024 Jul;67(1):129-133. [doi: 10.1016/j.amepre.2024.03.002] [Medline: 38479565]
3. Tsui FR, Shi L, Ruiz V, et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA Open 2021 Jan;4(1):ooab011. [doi: 10.1093/jamiaopen/ooab011] [Medline: 33758800]
4. Yoshimasu K, Kiyohara C, Miyashita K, Stress Research Group of the Japanese Society for Hygiene. Suicidal risk factors and completed suicide: meta-analyses based on psychological autopsy studies. Environ Health Prev Med 2008 Sep;13(5):243-256. [doi: 10.1007/s12199-008-0037-x] [Medline: 19568911]

5. Ribeiro JD, Franklin JC, Fox KR, et al. Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. Psychol Med 2016 Jan;46(2):225-236. [doi: 10.1017/S0033291715001804] [Medline: 26370729]
6. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. Ann Intern Med 2010 Nov 2;153(9):600-606. [doi: 10.7326/0003-4819-153-9-201011020-00010] [Medline: 21041580]
7. Practice Manual for Establishing and Maintaining Surveillance Systems for Suicide Attempts and Self-Harm: World Health Organization; 2018. URL: https://www.who.int/publications/i/item/practice-manual-for-establishing-and-maintaining-surveillance-systems-for-suicide-attempts-and-self-harm [accessed 2026-01-29]
8. Gahm GA, Reger MA, Kinn JT, Luxton DD, Skopp NA, Bush NE. Addressing the surveillance goal in the national strategy for suicide prevention: the department of defense suicide event report. Am J Public Health 2012 Mar;102(Suppl 1):S24-S28. [doi: 10.2105/AJPH.2011.300574] [Medline: 22390595]
9. Metzger MH, Tvardik N, Gicquel Q, Bouvry C, Poulet E, Potinet-Pagliaroli V. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study. Int J Methods Psychiatr Res 2017 Jun;26(2):e1522. [doi: 10.1002/mpr.1522] [Medline: 27634457]
10. Emergency Department Surveillance of Nonfatal Suicide-Related Outcomes (ED-SNRO). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/suicide/programs/ed-snsro.html [accessed 2022-1-27]
11. 2024 National Strategy for Suicide Prevention. Centers for Disease Control. 2024. URL: https://www.hhs.gov/programs/prevention-and-wellness/mental-health-substance-use-disorder/national-strategy-suicide-prevention/index.html [accessed 2026-01-27]
12. Bonilla AG, Pourat N, Chuang E, et al. Mental health staffing at HRSA-funded health centers may improve access to care. Psychiatr Serv 2021 Sep 1;72(9):1018-1025. [doi: 10.1176/appi.ps.202000337] [Medline: 34074146]
13. Gross TK, Lane NE, Timm NL, Committee on Pediatric Emergency Medicine. Crowding in the emergency department: challenges and best practices for the care of children. Pediatrics 2023 Mar 1;151(3):e2022060972. [doi: 10.1542/peds.2022-060972] [Medline: 36806666]
14. Roberts BK, Nofi CP, Cornell E, Kapoor S, Harrison L, Sathya C. Trends and disparities in firearm deaths among children. Pediatrics 2023 Sep 1;152(3):e2023061296. [doi: 10.1542/peds.2023-061296] [Medline: 37599647]
15. Naik-Mathuria BJ, Cain CM, Alore EA, Chen L, Pompeii LA. Defining the full spectrum of pediatric firearm injury and death in the United States: it is even worse than we think. Ann Surg 2023 Jul 1;278(1):10-16. [doi: 10.1097/SLA.0000000000005833] [Medline: 36825500]
16. Cantor J, Schuler MS, Kerber R, Purtle J, McBain RK. Changes in specialty crisis services offered before and after the launch of the 988 suicide and crisis lifeline. JAMA Psychiatry 2025 Apr 1;82(4):379-385. [doi: 10.1001/jamapsychiatry.2024.4548] [Medline: 39878975]
17. Boggs JM, Kafka JM. A critical review of text mining applications for suicide research. Curr Epidemiol Rep 2022;9(3):126-134. [doi: 10.1007/s40471-022-00293-w] [Medline: 35911089]
18. Zwald ML, Holland KM, Annor F, et al. Monitoring suicide-related events using National Syndromic Surveillance Program data. Online J Public Health Inform 2019;11(1):e440. [doi: 10.5210/ojphi.v11i1.9927]
19. Bey R, Cohen A, Trebossen V, et al. Natural language processing of multi-hospital electronic health records for public health surveillance of suicidality. NPJ Mental Health Res 2024;3(1):6. [doi: 10.1038/s44184-023-00046-7]
20. Edgcomb JB, Tseng CH, Pan M, Klomhaus A, Zima BT. Assessing detection of children with suicide-related emergencies: evaluation and development of computable phenotyping approaches. JMIR Ment Health 2023 Jul 21;10:e47084. [doi: 10.2196/47084] [Medline: 37477974]
21. Edgcomb JB, Olde Loohuis L, Tseng CH, et al. Electronic health record phenotyping of pediatric suicide-related emergency department visits. JAMA Netw Open 2024 Oct 1;7(10):e2442091. [doi: 10.1001/jamanetworkopen.2024.42091] [Medline: 39470636]
22. Rossom RC, Richards JE, Sterling S, et al. Connecting research and practice: implementation of suicide prevention strategies in learning health care systems. Psychiatr Serv 2022 Feb 1;73(2):219-222. [doi: 10.1176/appi.ps.202000596] [Medline: 34189931]
23. Oexle N, Feigelman W, Sheehan L. Perceived suicide stigma, secrecy about suicide loss and mental health outcomes. Death Stud 2020;44(4):248-255. [doi: 10.1080/07481187.2018.1539052] [Medline: 30451645]
24. Rockett IRH, Wang S, Stack S, et al. Race/ethnicity and potential suicide misclassification: window on a minority suicide paradox? BMC Psychiatry 2010 May 19;10(1):35. [doi: 10.1186/1471-244X-10-35] [Medline: 20482844]
25. Green C, Gottschlich EA, Burr WH. A national survey of pediatricians' experiences and practices with suicide prevention. Acad Pediatr 2023;23(7):1403-1410. [doi: 10.1016/j.acap.2023.01.010] [Medline: 36731651]
26. Horowitz LM, Bridge JA, Tipton MV, et al. Implementing suicide risk screening in a pediatric primary care setting: from research to practice. Acad Pediatr 2022 Mar;22(2):217-226. [doi: 10.1016/j.acap.2021.10.012] [Medline: 35248306]

27. Downs J, Velupillai S, George G, et al. Detection of suicidality in adolescents with autism spectrum disorders: developing a natural language processing approach for use in electronic health records. AMIA Annu Symp Proc 2018;2017:641-649. [Medline: 29854129]
28. Oquendo MA, Mann JJ. Suicidal behavior: a developmental perspective. Psychiatr Clin North Am 2008 Jun;31(2):xiii-xxvi. [doi: 10.1016/j.psc.2008.03.001] [Medline: 18439441]
29. Hedegaard H, Schoenbaum M, Claassen C, Crosby A, Holland K, Proescholdbell S. Issues in developing a surveillance case definition for nonfatal suicide attempt and intentional self-harm using International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) coded data. Natl Health Stat Report 2018 Feb(108):1-19. [Medline: 29616901]
30. Hoge MA, Vanderploeg J, Paris M, Lang JM, Olezeski C. Emergency department use by children and youth with mental health conditions: a health equity agenda. Community Ment Health J 2022 Oct;58(7):1225-1239. [doi: 10.1007/s10597-022-00937-7] [Medline: 35038073]
31. Obeid JS, Dahne J, Christensen S, et al. Identifying and predicting intentional self-harm in electronic health record clinical notes: deep learning approach. JMIR Med Inform 2020 Jul 30;8(7):e17784. [doi: 10.2196/17784] [Medline: 32729840]
32. Ji S, Pan S, Li X, Cambria E, Long G, Huang Z. Suicidal ideation detection: a review of machine learning methods and applications. IEEE Trans Comput Soc Syst 2021;8(1):214-226. [doi: 10.1109/TCSS.2020.3021467]
33. Adekkanattu P, Furmanchuk A, Wu Y, et al. Deep learning for identifying personal and family history of suicidal thoughts and behaviors from EHRs. NPJ Digit Med 2024 Sep 28;7(1):260. [doi: 10.1038/s41746-024-01266-7] [Medline: 39341983]
34. Bunnell BE, Tsalatsanis A, Chaphalkar C, et al. Automated detection and prediction of suicidal behavior from clinical notes using deep learning. PLoS ONE 2025;20(9):e0331459. [doi: 10.1371/journal.pone.0331459] [Medline: 40953025]
35. Martinez-Romo J, Araujo L, Reneses B. Guardian-BERT: early detection of self-injury and suicidal signs with language technologies in electronic health reports. Comput Biol Med 2025 Mar;186:109701. [doi: 10.1016/j.compbiomed.2025.109701] [Medline: 39967190]
36. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025 Jan 28;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
37. Zima BT, Gay JC, Rodean J, et al. Classification system for International Classification of Diseases, Ninth Revision, Clinical Modification and Tenth Revision pediatric mental health disorders. JAMA Pediatr 2020 Jun 1;174(6):620-622. [doi: 10.1001/jamapediatrics.2020.0037] [Medline: 32202603]
38. Posner K, Brown GK, Stanley B, et al. The Columbia-Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am J Psychiatry 2011 Dec;168(12):1266-1277. [doi: 10.1176/appi.ajp.2011.10111704] [Medline: 22193671]
39. Carson NJ, Mullin B, Sanchez MJ, et al. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. PLoS ONE 2019;14(2):e0211116. [doi: 10.1371/journal.pone.0211116] [Medline: 30779800]
40. Velupillai S, Epstein S, Bittar A, Stephenson T, Dutta R, Downs J. Identifying suicidal adolescents from mental health records using natural language processing. Stud Health Technol Inform 2019 Aug 21;264:413-417. [doi: 10.3233/SHTI190254] [Medline: 31437956]
41. Cusick M, Velupillai S, Downs J, et al. Portability of natural language processing methods to detect suicidality from clinical text in US and UK electronic health records. J Affect Disord Rep 2022 Dec;10:100430. [doi: 10.1016/j.jadr.2022.100430] [Medline: 36644339]
42. Coley RY, Johnson E, Simon GE, Cruz M, Shortreed SM. Racial/ethnic disparities in the performance of prediction models for death by suicide after mental health visits. JAMA Psychiatry 2021 Jul 1;78(7):726-734. [doi: 10.1001/jamapsychiatry.2021.0493] [Medline: 33909019]
43. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Int J Surg 2014 Dec;12(12):1495-1499. [doi: 10.1016/j.ijsu.2014.07.013] [Medline: 25046131]
44. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 16;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
45. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025 Jan;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
46. NOT-OD-15-089: racial and ethnic categories and definitions for NIH diversity programs and for other reporting purposes. National Institutes of Health. URL: https://grants.nih.gov/grants/guide/notice-files/not-od-15-089.html [accessed 2025-07-03]
47. Social vulnerability index. Agency for Toxic Substances and Disease Registry. 2020. URL: https://www.atsdr.cdc.gov/place-health/php/svi/index.html
48. Maroko AR, Doan TM, Arno PS, Hubel M, Yi S, Viola D. Integrating social determinants of health with treatment and prevention: a new tool to assess local area deprivation. Prev Chronic Dis 2016 Sep 15;13:E128. [doi: 10.5888/pcd13.160221] [Medline: 27634778]

49. Posner K, Oquendo MA, Gould M, Stanley B, Davies M. Columbia Classification Algorithm of Suicide Assessment (C-CASA): classification of suicidal events in the FDA's pediatric suicidal risk analysis of antidepressants. Am J Psychiatry 2007 Jul;164(7):1035-1043. [doi: 10.1176/ajp.2007.164.7.1035] [Medline: 17606655]
50. Zhang Y, Cai T, Yu S, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nat Protoc 2019 Dec;14(12):3426-3444. [doi: 10.1038/s41596-019-0227-6] [Medline: 31748751]
51. Aumüller M, Bernhardsson E, Faithfull A. ANN-Benchmarks: a benchmarking tool for approximate nearest neighbor algorithms. Inf Syst 2020 Jan;87:101374. [doi: 10.1016/j.is.2019.02.006]
52. Breiman L. Random forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
53. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. WIREs Data Min Knowl 2019 May;9(3):e1301. [doi: 10.1002/widm.1301]
54. Bayle P, Bayle A, Janson L, Mackey L. Cross-validation confidence intervals for test error. In: Proceedings of the 34th International Conference on Neural Information Processing Systems: Curran Associates Inc.; 2020:16339-16350 URL: https://dl.acm.org/doi/10.5555/3495724.3497095 [accessed 2025-12-13]
55. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837-845. [doi: 10.2307/2531595] [Medline: 3203132]
56. Kearns M, Neel S, Roth A, Wu ZS. Preventing fairness gerrymandering: auditing and learning for subgroup fairness. 2018 Presented at: Proceedings of the 35th International Conference on Machine Learning URL: https://proceedings.mlr.press/v80/kearns18a/kearns18a.pdf [accessed 2026-01-27]
57. Sveticic J, Stapelberg NCJ, Turner K. Suicidal and self-harm presentations to emergency departments: the challenges of identification through diagnostic codes and presenting complaints. Health Inf Manag 2020 Jan;49(1):38-46. [doi: 10.1177/1833358319857188] [Medline: 31272232]
58. Arias SA, Boudreaux ED, Chen E, et al. Which chart elements accurately identify emergency department visits for suicidal ideation or behavior? Arch Suicide Res 2019;23(3):382-390. [doi: 10.1080/13811118.2018.1472691] [Medline: 29791300]
59. Bommersbach TJ, McKean AJ, Olfson M, Rhee TG. National trends in mental health–related emergency department visits among youth, 2011-2020. JAMA 2023 May 2;329(17):1469-1477. [doi: 10.1001/jama.2023.4809] [Medline: 37129655]
60. Anderson KN, Johns D, Holland KM, et al. Emergency department visits involving mental health conditions, suicide-related behaviors, and drug overdoses among adolescents - United States, January 2019-February 2023. MMWR Morb Mortal Wkly Rep 2023 May 12;72(19):502-512. [doi: 10.15585/mmwr.mm7219a1] [Medline: 37167103]
61. Bertuccio P, Amerio A, Grande E, et al. Global trends in youth suicide from 1990 to 2020: an analysis of data from the WHO mortality database. EClinicalMedicine 2024 Apr;70:102506. [doi: 10.1016/j.eclinm.2024.102506] [Medline: 38440131]
62. Pompili M, Serafini G, Innamorati M, et al. Suicide risk in first episode psychosis: a selective review of the current literature. Schizophr Res 2011 Jun;129(1):1-11. [doi: 10.1016/j.schres.2011.03.008] [Medline: 21530179]
63. Kirby AV, Bakian AV, Zhang Y, Bilder DA, Keeshin BR, Coon H. A 20-year study of suicide death in a statewide autism population. Autism Res 2019 Apr;12(4):658-666. [doi: 10.1002/aur.2076] [Medline: 30663277]
64. Kõlves K, Fitzgerald C, Nordentoft M, Wood SJ, Erlangsen A. Assessment of suicidal behaviors among individuals with autism spectrum disorder in Denmark. JAMA Netw Open 2021 Jan 4;4(1):e2033565. [doi: 10.1001/jamanetworkopen.2020.33565] [Medline: 33433599]
65. Boggs JM, Yarborough BJH, Clarke G, et al. Development and validation of electronic health record measures of safety planning practices as part of zero suicide implementation. Arch Suicide Res 2025;29(3):654-667. [doi: 10.1080/13811118.2024.2394676] [Medline: 39193908]
66. Reyes-Portillo JA, Chin EM, Toso-Salman J, Blake Turner J, Vawdrey D, Mufson L. Using electronic health record alerts to increase safety planning with youth at-risk for suicide: a non-randomized trial. Child Youth Care Forum 2018 Jun;47(3):391-402. [doi: 10.1007/s10566-018-9435-4]
67. Bandealy A, Herrera N, Weissman M, Scheidt P. Use of lethal means restriction counseling for suicide prevention in pediatric primary care. Prev Med 2020 Jan;130:105855. [doi: 10.1016/j.ypmed.2019.105855] [Medline: 31644896]
68. Sisler SM, Hart S, Hamilton J, Schapiro NA. Preventing suicide through lethal means restriction in pediatric care. J Pediatr Health Care 2025;39(2):308-317. [doi: 10.1016/j.pedhc.2024.09.004] [Medline: 39797890]
69. Hoffmann JA, Stack AM, Monuteaux MC, Levin R, Lee LK. Factors associated with boarding and length of stay for pediatric mental health emergency visits. Am J Emerg Med 2019 Oct;37(10):1829-1835. [doi: 10.1016/j.ajem.2018.12.041] [Medline: 30600189]

Abbreviations:

aCS: augmented case surveillance
ADHD: attention-deficit or hyperactivity disorder
AUROC: area under the receiver operating characteristic curve
c-SSRS: Columbia Suicide Severity Rating Scale

CAMHD-CS: Child and Adolescent Mental Health Disorders Classification System
ED: emergency department
EHR: electronic health record
***ICD-10-CM*:** *International Classification of Diseases, 10th Revision, Clinical Modification*
ICD/CC: International Classification of Diseases codes and chief concerns
LLaMA: Large Language Model Meta AI
MH: mental health
NLP: natural language processing
ROC: receiver operating characteristic
SITB: self-injurious thoughts and behaviors
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by J Torous; submitted 13.Aug.2025; peer-reviewed by HH Yeh, J Obeid; revised version received 25.Nov.2025; accepted 28.Nov.2025; published 04.Feb.2026.

Please cite as:
Edgcomb JB, Saha A, Klomhaus A, Tascione E, Ponce CG, Lee JJ, Tacorda T, Zima BT
Detecting Pediatric Emergency Service Use for Suicide and Self-Harm: Multimodal Analysis of 3828 Encounters
JMIR Ment Health 2026;13:e82371
URL: https://mental.jmir.org/2026/1/e82371
doi:10.2196/82371

Original Paper

Evaluating a Culturally Tailored Digital Storytelling Intervention to Improve Trauma Awareness in Conflict-Affected Eastern Congo: Quasi-Experimental Pilot Study

Achille Bapolisi[1,2], MD, PhD; Jennifer Foucart[3], PhD; Déborah Kabambi[2], MD; Raïssa Mirishe[2], BSc; Elvis Musa[2], BSc; Aline Ruvunangiza[2], BSc; Joyce Bosomi[2], BSc; Victor Bulabula[2], MD; Marc Ilunga[4], LLD; Emmanuel Kajibwami[4], LLD; Odile Bapolisi[5], LLB; Arsene Daniel Nyalundja[6,7], MD; Marie-Hélène Igega[2], MD; Pacifique Mwene-batu[8], MD, PhD; Philippe de Timary[1], MD, PhD; Yasser Khazaal[9,10], MD

[1]Cliniques Universitaires Saint-Luc, Department of Psychiatry, Institute of Neurosciences, Faculté de Medecine, UCLouvain, Brussels, Brussels, Belgium

[2]Faculty of Medicine, Université Catholique de Bukavu, Bukavu, the Democratic Republic of the Congo

[3]Facultés des sciences de la motricité humaine, Université Libre de Bruxelles, Brussels, Belgium

[4]No academic affiliation at the time of publication, Bukavu, the Democratic Republic of the Congo

[5]Faculté de droit, Université Catholique de Bukavu, Bukavu, the Democratic Republic of the Congo

[6]Center for Tropical Diseases and Global Health, Université Catholique de Bukavu, Bukavu, the Democratic Republic of the Congo

[7]Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, United Kingdom

[8]Ecole Régionale de santé publique, Université Catholique de Bukavu, Bukavu, the Democratic Republic of the Congo

[9]Lausanne University Hospital, University of Lausanne, Lausanne, Switzerland

[10]Department of Psychiatry and Addictology, Université de Montréal, Montréal, QC, Canada

Corresponding Author:
Achille Bapolisi, MD, PhD
Cliniques Universitaires Saint-Luc, Department of Psychiatry, Institute of Neurosciences
Faculté de Medecine
UCLouvain
Avenue Hippocrate , 10
Brussels, Brussels
Belgium
Phone: 32 0466202635
Email: achilami@yahoo.fr

Abstract

Background: Posttraumatic stress disorder (PTSD) is highly prevalent in conflict-affected regions like eastern Democratic Republic of Congo; yet, cultural stigma and lack of psychoeducation limit public understanding and help-seeking behaviors.

Objective: This study evaluates the effect of a short, culturally adapted animated video on mental health perception, knowledge, and attitudes toward trauma.

Methods: A community-based quasi-experimental pre-post design was implemented among 239 participants from South Kivu. The intervention involved viewing a 3-minute animated psychoeducational video portraying locally relevant PTSD symptoms and resilience strategies. Perception, knowledge, and attitude scores were measured before and after the intervention, alongside PTSD prevalence and video appreciation.

Results: Out of 239, 40% (n=96) of the participants screened positively for PTSD. Post intervention, significant improvements were observed in perception (*P*=.01), knowledge (*P*<.001), and attitudes (*P*=.001) toward trauma. Appreciation was high; 82% (n= 195) expressed empathy for the characters, and 74% (n= 176) were likely to share the video. Linear regression showed that having PTSD symptoms (β coefficient=3.29, SE=1.09; *P*=.003), years of education (β coefficient=0.54, SE=0.08; *P*<.001), empathy toward the portrayed situations (β coefficient=5.07, SE=0.56; *P*<.001), perceived acquisition of new knowledge (β coefficient=2.58, SE=0.59; *P*<.001) and willingness to share the video (β coefficient=1.75, SE=0.50; *P*=.001) predicted stronger positive effect. A multiple linear regression including all predictors revealed that PTSD symptoms (β coefficient=1.93, SE=0.90; *P*=.03), years of education (β coefficient=0.47, SE=0.07; *P*<.001), empathy toward the portrayed situations (β coefficient=3.50, SE=0.55; *P*<.001), and willingness to share the video (β coefficient=1.75, SE=0.50; *P*=.001) remained significant predictors of video impact. Age and perceived acquisition of new knowledge were not significant in the multivariate model. This model accounted for 44.6% of the variance in video impact scores (R^2=0.446, $F_{6,231}$=30.99, *P*<.001).

Conclusions: This study highlights the effectiveness of culturally grounded, low-cost digital media for improving mental health literacy in postconflict settings. Video-based tools may serve as scalable components of trauma-informed care and public health communication in low-resource, high-need areas.

(JMIR Ment Health 2026;13:e81291) doi:10.2196/81291

KEYWORDS

post-traumatic stress disorder; digital media; mental health awareness; culturally adapted intervention; educational technology; psychoeducation; armed conflicts; mobile phone

Introduction

Raising awareness and educating the public about mental health is a major challenge, both in the fields of public health and psychotraumatology [1]. Better knowledge of mental disorders contributes not only to their prevention and treatment but also to the promotion of psychological well-being and the rehabilitation of those affected. Ideally, awareness strategies should rely on accessible, low-cost, widely shareable, and culturally appropriate means of communication.

Mental health disorders are increasingly recognized as a major contributor to morbidity and mortality worldwide, with an expected growing burden in low- and middle-income countries. Mental health conditions are among the top 10 causes of years lived with disability globally. In 2021, they made up around 15.6%-17.2% of all years lived with disability [2]. Among these disorders, posttraumatic stress disorder (PTSD) is drawing increasing attention from clinicians, researchers, and public health policymakers due to its rising prevalence, its psychological and relational impact, and its numerous psychiatric and somatic comorbidities [3]. The emergence of PTSD is closely linked to the growing frequency of traumatic events worldwide: armed conflicts, acts of terrorism, natural disasters exacerbated by climate change, road accidents, and more. One international study estimates that 70% of the global population has been exposed to at least 1 potentially traumatic event during their lifetime [4].

In response to this reality, movements such as trauma-informed interventions [5] and trauma-informed care [6] have emerged to promote greater recognition of trauma in both society and health care. These approaches aim to legitimize psychological suffering, raise awareness of trauma's impacts, and encourage clinical responses that are empathetic and appropriate. Despite these advances, many regions of the world, particularly in sub-Saharan Africa, remain largely unaffected by such awareness-raising initiatives and the integration of trauma into mental health policies [7]. Moreover, in these regions, mental health remains a marginal concern due to limited resource allocation and low levels of mental health literacy, both of which contribute to reduced accessibility and acceptability of mental health services [8].

Cultural differences largely explain the low levels of information, recognition, and acceptance of psychiatric care in these contexts [9]. This is especially concerning given that these same regions are often the most exposed to conflict. The Democratic Republic of Congo, marked by nearly 3 decades of armed violence resulting in more than 4 million deaths [10], is a stark illustration of this situation. In the most affected areas, PTSD prevalence reaches up to 40% [11]. The resurgence of conflict in February 2025 has further exacerbated the situation, profoundly impacting local populations. Yet, the mental health response is hindered by several obstacles, such as lack of information and awareness, stigma, reliance on alternative belief systems, and a shortage of specialized resources [12-14]. Patients with psychological disturbances often seek help from traditional healers before turning to psychologists or psychiatrists. Moreover, the historical divide between modern medical approaches and traditional practices has deepened a gap that remains difficult to bridge. Although the Democratic Republic of the Congo remains one of the poorest countries in the world [15], with over 52% of its population living on less than US $2.15 per day [16], it is nonetheless experiencing the effects of global digitalization, particularly in major urban areas where smartphone use and internet access are rapidly increasing [17]. This rapid digitalization offers new opportunities to extend communication and improve mental health literacy among the population. Recent digital health research underscores the importance of culturally sensitive design in improving user engagement and intervention outcomes across diverse settings [18].

To address these challenges, we developed an innovative approach that combines the opportunities offered by digitalization and the growing use of social media in the Democratic Republic of Congo with local cultural representations. We designed a short video featuring characters that portray common manifestations of PTSD while suggesting possible paths to resilience. Short video–based interventions have recently shown strong potential to improve mental health literacy and reduce stigma in community settings [19]. Digital storytelling was chosen over other digital interventions because it combines narrative, emotion, and cultural resonance—3 elements shown to strengthen engagement and knowledge retention in health communication. In eastern Congo, where literacy levels vary widely, and stigma surrounding mental illness remains strong, stories conveyed through relatable characters and familiar cultural symbols can communicate complex psychological concepts more effectively than text-based or purely informational approaches. The narrative format allows individuals to recognize their own experiences within the story, promoting empathy, normalization of psychological distress, and reduction of stigma. Moreover, the use of animation and local languages increases accessibility across literacy levels and cultural groups. This approach aligns with growing evidence that culturally grounded digital storytelling can enhance mental health literacy and foster behavioral intentions to seek help in low-resource and conflict-affected contexts [20-22].

This pilot study was designed to explore the potential of such a video-based intervention to enhance mental health perceptions, knowledge, and attitudes in South Kivu, a region heavily affected by armed violence. This quasi-experimental pilot study was conducted in South Kivu, a region heavily affected by armed conflict, to assess the intervention's potential to improve trauma-related perceptions, knowledge, and attitudes. The study also sought to provide preliminary insights into the acceptability, feasibility, and short-term impact of this low-cost, scalable psychoeducational tool. We hypothesized that viewing the culturally adapted video would result in (1) improved perceptions and knowledge about psychological trauma and (2) more positive attitudes toward professional help-seeking and social support compared with preintervention measures.

Methods

Video Capsule Development Process

An initial team composed of YK, JF, PdT, and AB outlined the main features of the video capsule. It was designed to be no longer than 3 minutes and aimed primarily to provide trauma-focused psychoeducation in a culturally adapted format. The duration was deliberately limited to less than 3 minutes, as shorter videos have been shown to maximize viewer engagement, comprehension, and completion rates in health education contexts, particularly among audiences with varying literacy levels [23-27]. In addition, concise videos are easier to share through popular social media and mobile messaging platforms commonly used in the Democratic Republic of Congo, facilitating wider community dissemination [17].

AB subsequently conducted interviews with 1 psychiatrist and 2 psychologists with extensive clinical experience in the region, to identify the most commonly reported symptoms and the most frequently mobilized resilience strategies. Reported symptoms included nightmares, flashbacks, sadness, palpitations, and difficulties with concentration. Additionally, some interviews highlighted a local representation of trauma as a form of spiritual attack or curse affecting several members within a community.

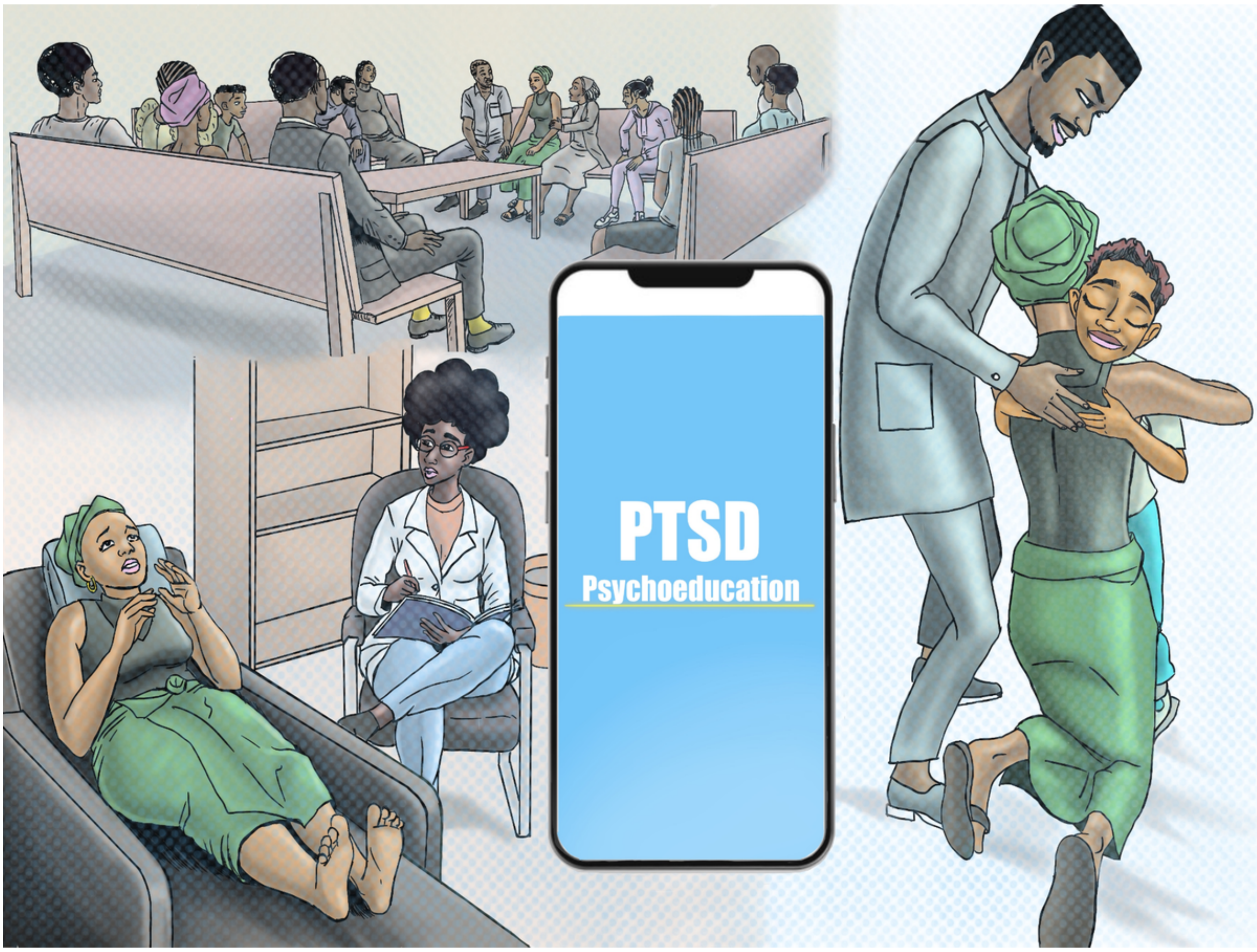
In terms of resilience strategies, community and family support, physical activity, and psychomedical follow-up were cited as the most helpful resources. Based on this input, AB drafted an initial script featuring a diverse set of characters intended to reflect the pluralistic nature of society.

This script was shared with local illustrators (MI and EK) familiar with cultural and contextual specificities. A collaborative staging process followed, allowing for detailed discussions of each visual and narrative element (Multimedia Appendix 1).

Once the final version of the video capsule was completed, AB presented it to various groups, including students (n=50), merchants (n=29), and colleagues in psychiatry and psychology (n=5), to assess its clarity, intelligibility, and relevance. To enhance accessibility, the video used simplified language and clear visual storytelling, allowing key messages to be understood without reliance on text. Narration was provided in 5 local languages—Swahili, Lingala, Tshiluba, Kikongo, and French—to support comprehension among participants with low literacy or mild communication difficulties. These adaptations ensured inclusivity and cultural relevance across diverse audiences, and versions with on-screen captions are currently being developed to further improve accessibility.

Some illustrative scenes from the culturally adapted animated video used in the intervention are presented in Figure 1. The animation portrays locally recognizable experiences and emotional responses to trauma, including fear, distress, and social withdrawal, as well as supportive interactions with health care professionals and community members. Characters were designed to reflect the cultural and social diversity of eastern Congo, aiming to foster empathy, identification, and understanding of PTSD and pathways to resilience.

Figure 1. Illustrative scenes from the culturally adapted animated video used in the intervention.



Study Design and Setting

We conducted a community-based quasi-experimental pre-post study to evaluate changes in perception, knowledge, and attitude related to mental health following the viewing of an educational video capsule.

Measurements were taken immediately before and after the intervention. The study used an immediate postintervention assessment to capture the short-term effects of the video on perceptions, knowledge, and attitudes toward psychological trauma. This approach was chosen for both methodological and contextual reasons. Given the security volatility in eastern Congo, the risk of population displacement, and limited communication infrastructure, a follow-up assessment was not feasible within the study's timeframe. Immediate postassessment ensured that the same participants could be evaluated under consistent conditions while minimizing attrition bias. Moreover, as a pilot study, the primary objective was to determine the intervention's short-term acceptability and feasibility rather than long-term behavioral change. Nevertheless, we acknowledge that perceptions, knowledge, and attitudes are constructs that influence behavioral intentions and actions over time, and future studies are planned to include longitudinal follow-ups to assess retention and potential behavioral outcomes.

The study took place in the eastern region of the Democratic Republic of Congo, specifically in the city of Bukavu and the rural surrounding areas of Nyantende, Idjwi, Ciriri, and Kalehe. These regions have historically been affected by repeated episodes of armed conflict. Data collection was conducted between December 2024 and January 2025, during a period of relative calm that preceded the escalation of conflict in February 2025.

Inclusion and Exclusion Criteria

All adults aged 18 years and older who had been residing in Bukavu, Idjwi, Kalehe, Ciriri, or Nyatende for at least 6 months were eligible to participate.

Individuals who were severely ill, exhibited severe mental disorders, or had significant communication impairments (eg, hearing or speech difficulties and inability to provide informed consent) were excluded from the study.

Sample Size Determination

The sample size was calculated to detect a statistically significant change in perception of psychotrauma, measured as a continuous variable, using a paired-sample *t* test. With a 2-tailed α of .05, 80% power, a moderate effect size (Cohen *d*=0.5), and assuming a pre-post correlation of 0.5, the required minimum sample size was estimated at 34 participants, using G*Power software. However, our final sample consisted of 239 participants, which not only met the requirement for the primary outcome but also allowed for broader analyses, including the estimation of the prevalence of PTSD.

Sampling Procedure

A multistage sampling technique was used to ensure both representativeness and logistical feasibility. South Kivu province is administratively divided into several territories and urban communes that vary substantially in population density and exposure to conflict. In the first stage, 5 zones were selected at random using a lottery method from the list of health zones in the province. In the second stage, within each selected zone, districts and rural sectors were stratified according to their population size and accessibility. From these, 1 urban district and 4 rural sectors were randomly selected.

Within each selected area, approximately 30% of neighborhoods were then randomly chosen. This proportion was determined a priori as a balance between representativeness and operational feasibility, given constraints related to time, security, and available field personnel. The 30% threshold was therefore fixed across sites to maintain comparability of sampling intensity, while the specific neighborhoods were selected randomly within each site based on administrative population listings provided by local authorities.

Finally, within each selected neighborhood, the target number of participants was proportionally allocated according to estimated population size, and households were approached consecutively until quotas were met. Because this was an exploratory pilot study rather than a population-weighted survey, clustering effects were not incorporated in the sample size calculation. However, data collection procedures were standardized across sites to minimize intercluster variability, and potential design effects were evaluated during analysis through comparison of site-level means and variances. No significant clustering effect was observed. The intervention was delivered individually to each participant using smartphones operated by trained research assistants. Participants viewed the video once in a quiet setting immediately before completing the postintervention assessment. This personalized format was chosen to ensure that all participants could clearly see and hear the content. All research assistants followed a standardized protocol to maintain fidelity across sites, including identical video files, playback settings, and instructions. Data collection procedures were harmonized through a detailed manual and a 3-day training workshop focused on consistent delivery, participant guidance, and confidentiality. Site-level consistency was verified during analysis, and no significant variation in intervention effects was observed between locations, in coherence with the standardization of delivery and data collection procedures.

Perception, Knowledge, and Attitudes Toward Trauma

The perception questionnaire was adapted from the studies on perceptions in mental health [28], and included 4 items. The knowledge and attitude sections were adapted from the study on trauma-informed care [29], comprising 7 items on knowledge and 6 items on attitude.

Perception was assessed using an inverse Likert scale ranging from 0 (strongly agree) to 4 (strongly disagree), while knowledge and attitudes were assessed using a standard Likert scale ranging from 0 (strongly disagree) to 4 (strongly agree). Individual scores were summed to produce total scores for each domain.

To ensure cultural and contextual accuracy, the questionnaires were translated into French and Swahili, then independently back-translated into English by bilingual experts. The translated versions were reviewed by a panel of local psychiatrists and psychologists to verify conceptual equivalence and cultural appropriateness of terms and examples. Before data collection, the instruments were pilot tested among 20 adults from Bukavu and Idjwi to assess clarity, comprehensibility, and cultural relevance. Feedback from this pilot phase led to minor wording adjustments to reflect local idioms of distress and expressions of emotional states. Internal consistency reliability for the final scales was satisfactory (Cronbach α=0.81 for perception, 0.86 for knowledge, and 0.84 for attitudes).

The effect score of the video was calculated as the difference between the total score obtained after viewing and the total score obtained before viewing, by summing all item scores at each time point.

Appreciation of the Video Capsule

We also assessed participants' appreciation of the video capsule using 3 questions that evaluated (1) empathy toward the characters and situations portrayed, (2) perceived acquisition of new knowledge, and (3) willingness to share the video. Responses were measured on a Likert scale ranging from *not at all* to *extremely*. We calculated the number and percentage of participants selecting each response level.

Psychological Assessment

We measured past traumatic events and PTSD through the Post-traumatic Diagnostic Scale - French adaptation (PDS-F), along with a Stressful Events Scale [30], a detailed scale assessing the types and magnitude of a wide variety of traumatic events as well as PTSD symptoms. This scale showed good psychometric values in African populations [31] and has a validated French version [30]. The PDS-F is interpreted with a severity score ranging from 0 to 51 obtained by adding up the responses of items. For the positive screening of PTSD to be made, we considered the cutoff for moderate to severe symptoms, a rating of >20 [32].

Data Collection Procedures and Quality Control

Data were collected by 5 medical students. The principal investigator (AB) trained the medical students over 3 days, emphasizing the theoretical and practical aspects of all questions in the questionnaire, informed consent, and participant confidentiality.

Ethical Considerations

Ethical approval was obtained from the Catholic University of Bukavu Ethics committee (UCB/CIES/NC/033/253). We obtained written informed consent from participants, and we ensured their privacy and confidentiality. Participants were then offered, after the completion of the evaluation sessions, with the main investigator, where needed, and those with psychological disturbances were advised to attend counseling sessions in the psychiatric clinic of the hospital. No financial or material compensation was provided for participation.

Data Analysis

We analyzed data using Stata (version 13; StataCorp) to perform descriptive and inferential analysis. Qualitative variables were described in terms of frequencies and percentages, while continuous variables were described in terms of means and SDs. Pearson chi-square and the Student *t* test were used to compare characteristics of patients for categorical and continuous variables, respectively.

We conducted bivariate and multiple linear regression analyses to explore associations between demographic, psychological, and appreciation-related variables and the overall effect score on perceptions, knowledge, and attitudes toward trauma. Variables showing a bivariate association with the outcome at $P\leq.05$ were entered into the multivariate model. Model building followed a stepwise strategy: theoretically relevant variables (age, sex, education level, and PTSD symptom severity) were entered first to control for potential confounding, followed by appreciation variables (empathy, perceived acquisition of new knowledge, and willingness to share the video) reflecting emotional engagement and behavioral intention. The inclusion of these predictors was guided by the theory of planned behavior and previous evidence linking affective engagement and trauma exposure to psychoeducational outcomes. Model fit and multicollinearity were assessed using the coefficient of determination (R^2), adjusted R^2, *F* statistics, and variance inflation factors (VIFs; all VIF<2), confirming appropriate specification and absence of collinearity.

We conducted bivariate and multiple linear regression analyses to explore associations between demographic, psychological, and appreciation-related variables and the overall effect score on perceptions, knowledge, and attitudes toward trauma. Variables showing a bivariate association with the outcome at $P\leq.05$ were entered into the multivariate model. Model building followed a stepwise strategy: theoretically relevant variables (age, sex, education level, and PTSD symptom severity) were entered first to control for potential confounding, followed by appreciation variables (empathy, perceived acquisition of new knowledge, and willingness to share the video) reflecting emotional engagement and behavioral intention. The inclusion of these predictors was guided by the theory of planned behavior and previous evidence linking affective engagement and trauma exposure to psychoeducational outcomes. Empathy enhances perspective taking and prosocial intentions, which are associated with greater openness toward trauma-related information [33-37]. Likewise, perceived self-relevance and social relevance are robust predictors of video sharing intentions, as individuals are more likely to share content they find personally meaningful or beneficial to their social network [38,39]. Emotional resonance—such as empathy, hope, or identification with the scenario—fosters authenticity and social connection, motivating sharing behavior and reinforcing learning through peer support [40]. Trust in the source or creator also predicts sharing behavior, with higher information and science literacy associated with more discerning and health-promoting dissemination [41,42]. In line with the theory of planned behavior, knowledge contributes to behavioral change by shaping attitudes, perceived behavioral control, and subjective norms, which in turn influence intentions and actions [42,43]. The inclusion of age, sex, and education level as predictors is further supported by extensive evidence showing their association with posttraumatic stress symptom severity and psychoeducational outcomes [44-46]. Model fit and multicollinearity were assessed using R^2, adjusted R^2, *F* statistics, and VIFs (all VIF<2), confirming appropriate specification and absence of collinearity.

Results

Sociodemographic Characteristics and PTSD Prevalence

A total of 239 participants were enrolled, comprising 133 (56%) women and 106 (44%) men, with a mean age of 34 (SD 14) years. Of the total, 147 (62%) participants were married, and 79 (33%) were unemployed. Regarding education, 86 (36%) had completed university education, and the mean number of years of schooling was 10 (SD 7) years. There were significant sex differences in marital status, profession, and education level ($P<.05$). The overall prevalence of probable PTSD (positive screening for PTSD) was 40%, with no significant difference between sex ($P=.22$; Table 1).

Table 1. Sociodemographic characteristics and posttraumatic stress disorder prevalence (N=239).

Variable	Total	Female	Male	*P* value
Sex, n (%)	239 (100)	133 (56)	106 (44)	—[a]
Age (y), mean (SD)	34 (14)	34 (13)	34 (14)	.92
Marital status, n (%)				.002
Married	147 (62)	91 (68)	56 (53)	
Single	78 (33)	31 (23)	47 (44)	
Separated or divorced	6 (2)	6 (5)	0 (0)	
Widower	8 (3)	5 (4)	3 (3)	
Occupation, n (%)				
Liberal activity	52 (22)	33 (25)	19 (18)	<.001
State official	19 (8)	5 (4)	14 (13)	.29
Humanitarian worker	13 (5)	3 (2)	10 (9)	.005
Unemployed	79 (33)	48 (36)	31 (29)	<.001
Student	32 (13)	11 (8)	21 (20)	<.001
Others	52 (23)	33 (25)	19 (18)	.21
Education level, n (%)				
Less than primary	46 (19)	36 (27)	10 (9)	—
Primary	30 (13)	23 (17)	7 (7)	.001
Secondary	77 (32)	45 (34)	32 (30)	.06
University	86 (36)	29 (22)	57 (54)	.29
Years of education, mean (SD)	10 (7)	8 (7)	12 (6)	<.001
PTSD[b] prevalence, n (%)	96 (40)	58 (44)	38 (36)	.22

[a]Not applicable.

[b]PTSD: posttraumatic stress disorder.

Perception, Knowledge, and Attitudes Toward Trauma

Following the video intervention, participants showed significant improvements in perception, knowledge, and attitude scores, respectively (Table 2). The mean perception score increased from 11.8 (SD 2.5) to 12.4 (SD 2.4; *P*=.01). Several perception items, such as the belief that "talking about one's suffering is useless," improved significantly (*P*=.01).

Table 2. Participant responses to video content (N=239).

Item	Likert scale response, n (%)				
	Not at all	Very little	Moderately	A lot	Extremely
To what extent did you feel empathy toward the people or situations portrayed in the video?	4 (1.7)	8 (3.3)	32 (13.4)	105 (43.9)	90 (37.7)
To what extent did the video provide you with new information or knowledge?	5 (2.1)	10 (4.2)	32 (13.4)	99 (41.4)	93 (38.9)
How likely are you to share this video with others in your network?	8 (3.3)	9 (3.8)	46 (19.2)	90 (37.7)	86 (36)

Knowledge scores improved markedly, rising from 19.3 (SD 5.5) to 21.2 (SD 3.9; *P*<.001). Significant changes were observed in items relating to trauma's effect on mental and physical health and recognition of trauma symptoms (eg, nightmares and palpitations).

Similarly, attitudes toward trauma improved, with the total score increasing from 16.2 (SD 4.3) to 17.5 (SD 4.7; *P*=.001). Key items, such as belief in the possibility of recovery and the need for professional support, showed significant positive shifts.

Effect sizes were calculated for each domain to complement significance testing. The intervention produced small to moderate improvements in perception (Cohen *d*=0.26), knowledge (Cohen *d*=0.38), and attitude (Cohen *d*=0.28). Site-level analyses were also performed to examine potential regional variation in outcomes; no significant differences were observed across sites, suggesting consistent intervention effects throughout the different study areas.

A detailed version of these results, including all item-level data, can be consulted in **Multimedia Appendix 2**.

Appreciation of the Video Capsule

Participants' feedback on the video was positive (Table 2). Out of 239 participants, 195 (82%) felt empathy "a lot" or "extremely" toward the portrayed situations, 192 (80%) reported having acquired new knowledge, and 176 (74%) stated that they were likely or very likely to share the video.

Predictors of Video Effect

Simple linear regression analyses revealed several significant predictors of higher video effect scores. Strongest associations were observed for video appreciations: higher levels of empathy toward the portrayed situations (β coefficient=5.07, SE=0.56, t_{237}=9.07; P<.001), perceived acquisition of new knowledge (β coefficient=2.58, SE=0.59, t_{237}=4.36; P<.001), and willingness to share the video (β coefficient=3.30, SE=0.53, t_{237}=6.23; P<.001) were all significantly associated with greater video effect. Among demographic and clinical variables, years of education showed a strong positive effect (β coefficient=0.54, SE=0.08, 95% CI 0.39-0.69, R^2=0.171; P<.001), while PTSD symptoms were also positively associated with video effect (β coefficient=3.29, SE=1.09, 95% CI 1.13-5.44, R^2=0.037; P=.003). In contrast, age was inversely associated with video effect (β coefficient=–0.11, SE=0.04, 95% CI –0.19 to –0.03, R^2=0.031; P=.007).

A multiple linear regression including all predictors revealed that PTSD symptoms (β coefficient=1.93, SE=0.90, t_{231}=2.14; P=.03), years of education (β coefficient=0.47, SE=0.07, t_{231}=6.69; P<.001), empathy toward the portrayed situations (β coefficient=3.50, SE=0.55, t_{231}=6.33; P<.001), and willingness to share the video (β coefficient=1.75, SE=0.50, t_{231}=3.47; P=.001) remained significant predictors of video effect. Age and perceived acquisition of new knowledge were not significant in the multivariate model. This model accounted for 44.6% of the variance in video effect scores (R^2=0.446, $F_{6,231}$=30.99; P<.001; Table 3).

Table 3. Multiple linear regression predicting video effect (N=238). Model fit: F6,231=30.99, R2=0.446, adjusted R2=0.432; root-mean-square error=6.60; P<.001.

Variable	β coefficient (SE)	95% CI	*t* test (*df*)	*P* value
Age	–0.003 (0.033)	−0.07 to 0.06	–0.10 (231)	.92
Years of education	0.47 (0.07)	0.33 to 0.61	6.69 (231)	<.001
PTSD[a] score	1.93 (0.90)	0.15 to 3.71	2.14 (231)	.03
Empathy toward the portrayed situations	3.50 (0.55)	2.41 to 4.59	6.33 (231)	<.001
Perceived acquisition of new knowledge	0.70 (0.55)	−0.38 to 1.78	1.28 (231)	.20
Willingness to share the video	1.75 (0.50)	0.76 to 2.74	3.47 (231)	.001
Constant	27.35 (2.55)	22.32 to 32.38	10.72 (231)	<.001

[a]PTSD: posttraumatic stress disorder.

Discussion

Principal Findings

This study demonstrated that a short, culturally adapted animated video significantly improved participants' perceptions, knowledge, and attitudes toward psychological trauma in conflict-affected eastern Congo. After viewing the video, mean scores for perception, knowledge, and attitudes increased markedly, with large effect sizes and high participant engagement—over 80% of viewers reported empathy for the characters and 74% indicated willingness to share the video. Regression analyses further revealed that higher education, empathy toward the portrayed situations, and the presence of PTSD symptoms were significant predictors of stronger intervention effects. Together, these findings indicate that culturally grounded, low-cost digital storytelling can effectively enhance mental health literacy and reduce stigma in low-resource, trauma-affected settings.

Our findings align with and extend evidence from video-based and multimedia psychoeducational interventions. For example, the effect of online multimedia psychoeducational interventions on the resilience and perceived stress of hospitalized patients with COVID-19 found significant improvements in resilience and reductions in perceived stress after a brief online multimedia intervention [47]. Our findings are also consistent with evidence from trauma-specific psychoeducational interventions, which similarly produced moderate improvements in PTSD, depression, and bonding outcomes among pregnant women with trauma histories, demonstrating the potential of brief, structured psychoeducation to enhance psychological well-being in vulnerable populations [48]. In our study, a culturally tailored, 3-minute animated video delivered individually, the small-to-moderate effect sizes we observed suggest that even very brief interventions can produce meaningful changes in mental-health literacy in conflict-affected settings.

The mechanisms underlying the observed changes likely stem from enhanced cognitive and emotional engagement achieved through a culturally grounded narrative. By portraying locally recognizable experiences and characters, the video fostered empathy, identification, and reflection—key processes known to mediate learning and attitude change in health communication [49,50]. Delivering the intervention individually via smartphones, with multilingual narration and culturally adapted visuals, further improved accessibility, attention, and comprehension, thereby supporting effective knowledge transfer and stigma reduction.

Low-cost, scalable digital interventions are increasingly recognized as practical strategies for improving health literacy and reducing stigma in resource-limited settings [51]. In this study, the creation of an animated video on PTSD, based on symptoms and experiences most frequently described by Congolese clinicians and patients, offered a familiar and emotionally engaging medium for psychoeducation. The animation, developed by local artists, embodied the region's cultural and linguistic diversity and avoided overly medicalized or Western frameworks that may alienate local audiences [8,52,53]. Such a design, consistent with human factors principles, enhances usability, accessibility, and long-term engagement with digital mental health content, even in low-resource contexts [54].

The changes observed in perception, knowledge, and attitudes suggest that video capsules like this could be an effective tool for psychoeducation and for reducing stigma associated with mental health issues. Previous studies have shown that individuals suffering from psychological disorders are often heavily stigmatized in these populations, where seeking mental health care is perceived as shameful [55,56]. Survivors of rape and torture are also frequently stigmatized due to perceived social shame and dishonor [57,58]. The video was therefore designed to legitimize the suffering resulting from trauma and to encourage understanding and social support.

One justified concern in war-affected contexts is the risk of retraumatization or vicarious trauma [6,59,60]. Empathy elicited by trauma-related video content is consistently associated with increased emotional distress, including heightened anxiety, depressive symptoms, and physiological stress responses, such as elevated heart rate and cortisol, particularly in individuals with high emotional contagion or affective empathy [34,37,61-63]. This issue was central to the development of the video, and efforts were made to avoid overly distressing images. The emotional impact of the content was carefully reviewed throughout the process. In addition, as described in our methodology, ethical safeguards were in place to recommend appropriate care for participants exhibiting signs of trauma. In addition to illustrating the psychological consequences of trauma, the video also depicts multiple pathways to resilience—individual coping efforts, family and community support, and access to psychomedical assistance—reflecting an integrated, culturally grounded approach to recovery.

Our findings also show that participants with PTSD reported higher empathy toward the characters and experienced a more pronounced positive impact on their perceptions, knowledge, and attitudes after viewing the video. One of the most striking findings of our study is the relatively high prevalence of positive screening for PTSD (40%). While this figure may appear elevated compared with other studies [64-66], it aligns closely with previous research conducted in the conflict-affected regions of eastern Congo [11]. Furthermore, the recent deterioration in regional security—marked by the resurgence of a rebel group that ultimately forced us to halt data collection—likely heightened the expression of traumatic symptoms within the population.

Finally, our study shows that the video's positive effect was particularly evident among young, educated individuals and those showing posttraumatic symptoms. This finding supports the targeted use of such interventions among youth, students, and individuals at risk of PTSD. These results point to the potential for implementing video-based awareness tools in schools, universities, and health care settings, contributing meaningfully to trauma-informed interventions and trauma-informed care, with both clinical and public health implications.

Taken together, our findings suggest that culturally contextualized video capsules can foster positive changes in how psychological trauma is perceived, understood, and addressed in war-affected regions. Our findings align with the theory of planned behavior [67], which posits that intention is a proximal predictor of behavior. The positive shifts in attitudes toward professional help, reduced stigma, and increased endorsement of social support suggest an enhanced behavioral intention to seek care. While actual care-seeking behavior was not directly measured, our results imply a potential increase in future help-seeking behaviors.

Limitations

While this study provides valuable insights into the potential of culturally adapted video interventions to improve mental health awareness in conflict-affected regions, several limitations should be acknowledged. First, the quasi-experimental pre-post design without a control group limits causal inference. Although improvements in perception, knowledge, and attitudes were observed after viewing the video, we cannot fully rule out the influence of external factors or testing effects. Second, the follow-up was immediate and did not assess the sustainability of changes over time. Longer-term follow-up would be necessary to determine whether improvements in knowledge and attitudes persist and translate into behavioral change or increased care-seeking. Third, the self-report nature of the measures may have introduced bias, including social desirability effects, particularly in postintervention responses. Additionally, while the instruments used were adapted from validated tools, further validation in this specific cultural context would strengthen the robustness of the findings. Although internal consistency coefficients were good, these indices primarily reflect internal reliability rather than the full spectrum of psychometric validity. Future research should therefore include confirmatory factor analyses and convergent or discriminant validity assessments to ensure that the constructs measured retain their theoretical structure and meaning in this cultural setting. Fourth, although efforts were made to ensure linguistic and cultural appropriateness, the diverse ethnic and linguistic backgrounds of participants in South Kivu mean that certain nuances may not have been fully captured by a single video version. Future studies could explore more tailored adaptations by subgroup and by level of instruction. Finally, the study sample, although geographically diverse, may not be representative of the broader Congolese population or other conflict-affected regions. Moreover, individuals with severe mental illness or communication impairments were excluded, possibly underestimating the intervention's reach and effect among the most vulnerable populations.

Conclusion

This study highlights the potential of culturally adapted, multimedia interventions to improve mental health literacy in conflict-affected regions. By integrating local representations of trauma with evidence-based psychoeducation, the video capsule successfully enhanced participants' perceptions, knowledge, and attitudes toward psychological trauma. The positive reception of the intervention, particularly among individuals with PTSD symptoms and those with higher education levels, underscores the value of contextually sensitive and accessible approaches to mental health promotion. In settings like eastern Democratic Republic of Congo, where traditional stigma, limited resources, and ongoing violence hinder mental health care, such tools may represent a promising, scalable strategy to foster trauma-informed awareness, reduce stigma, and support pathways to resilience. Continued investment in culturally grounded public health communication and further evaluation in large-scale and long-term impact will be essential for strengthening mental health systems in similar contexts.

Acknowledgments

We warmly thank Professor Ghislain Bisimwa, Professor Guy Mulinganya, Dr Ghislain Maheshe, and Professor Marius Baguma for their invaluable institutional support at the Université Catholique de Bukavu. We also extend our sincere gratitude to fUNIKIN-BEBUG and ASJ, a network of researchers in the Democratic Republic of Congo, for their guidance and support in the development of the study protocol.

The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to participant confidentiality, but are available from the corresponding author upon reasonable request.

Funding

This study was funded by the European Civis University Alliance as part of a Fonds de Partenariat privilégié UNIL ULB project between the University of Lausanne and the Université libre de Bruxelles.

Authors' Contributions

AB, JF, PdT, and YK designed the study, contributed to the video creation, conducted data collection and analysis, and prepared the first and revised drafts of the manuscript. DK, RM, AM, EM, OB, ADN, MHI, PM, VB, and EK contributed to the video creation as well as the preparation and editing of the manuscript. All authors have read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Sample video capsule (English version) used in the intervention. The 3-minute animated clip illustrates locally relevant manifestations of psychological trauma (such as nightmares, avoidance, and emotional distress) and highlights culturally meaningful resilience strategies, including community and family support, physical activity, and professional help-seeking. The video was produced collaboratively with Congolese illustrators and mental health professionals to ensure cultural and linguistic appropriateness.

[MP4 File (MP4 Video), 39314 KB - mental_v13i1e81291_app1.mp4]

Multimedia Appendix 2

UntitlTable : Perception, Knowledge, and Attitude Scores Before and After Visualizationed.

[DOC File , 29 KB - mental_v13i1e81291_app2.doc]

References

1. Moitra M, Owens S, Hailemariam M, Wilson KS, Mensa-Kwao A, Gonese G, et al. Global mental health: where we are and where we are going. Curr Psychiatry Rep 2023;25(7):301-311 [FREE Full text] [doi: 10.1007/s11920-023-01426-8] [Medline: 37256471]

2. World mental health report: transforming mental health for all. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240049338 [accessed 2025-01-15]
3. Atwoli L, Stein DJ, Koenen KC, McLaughlin KA. Epidemiology of posttraumatic stress disorder: prevalence, correlates and consequences. Curr Opin Psychiatry 2015;28(4):307-311 [FREE Full text] [doi: 10.1097/YCO.0000000000000167] [Medline: 26001922]
4. Benjet C, Bromet E, Karam EG, Kessler RC, McLaughlin KA, Ruscio AM, et al. The epidemiology of traumatic event exposure worldwide: results from the World Mental Health Survey Consortium. Psychol Med 2016;46(2):327-343 [FREE Full text] [doi: 10.1017/S0033291715001981] [Medline: 26511595]
5. Han H, Miller HN, Nkimbeng M, Budhathoki C, Mikhael T, Rivers E, et al. Trauma informed interventions: a systematic review. PLoS One 2021;16(6):e0252747 [FREE Full text] [doi: 10.1371/journal.pone.0252747] [Medline: 34157025]
6. Grossman S, Cooper Z, Buxton H, Hendrickson S, Lewis-O'Connor A, Stevens J, et al. Trauma-informed care: recognizing and resisting re-traumatization in health care. Trauma Surg Acute Care Open 2021;6(1):e000815 [FREE Full text] [doi: 10.1136/tsaco-2021-000815] [Medline: 34993351]
7. Johnson NE, Malimabe M, Yoon GH, Osborn TL, Falgas-Bague I, Swahn MH. Integration of local realities to address mental health in Africa. Lancet Psychiatry 2025;12(6):399-401. [doi: 10.1016/S2215-0366(25)00005-7] [Medline: 39986289]
8. Atewologun F, Adigun OA, Okesanya OJ, Hassan HK, Olabode ON, Micheal AS, et al. A comprehensive review of mental health services across selected countries in sub-Saharan Africa: assessing progress, challenges, and future direction. Discov Ment Health 2025;5(1):49 [FREE Full text] [doi: 10.1007/s44192-025-00177-7] [Medline: 40195169]
9. Wondimagegn D, Pain C, Seifu N, Cartmill C, Alemu AA, Whitehead CR. Reimagining global mental health in Africa. BMJ Glob Health 2023;8(9):e013232 [FREE Full text] [doi: 10.1136/bmjgh-2023-013232] [Medline: 37666576]
10. Mortality in the Democratic Republic of Congo: an ongoing crisis. International Rescue Committee. 2018. URL: https://www.rescue.org/report/mortality-democratic-republic-congo-ongoing-crisis [accessed 2026-01-06]
11. Johnson K, Scott J, Rughita B, Kisielewski M, Asher J, Ong R, et al. Association of sexual violence and human rights violations with physical and mental health in territories of the Eastern Democratic Republic of the Congo. JAMA 2010;304(5):553-562. [doi: 10.1001/jama.2010.1086] [Medline: 20682935]
12. On'okoko MO, Jenkins R, Miezi SMM, Andjafono DOLE, Mushidi IM. Mental health in the Democratic Republic of Congo: a post-crisis country challenge. Int Psychiatry 2010;7(2):41-42 [FREE Full text] [Medline: 31508032]
13. Saade S, Parent-Lamarche A, Khalaf T, Makke S, Legg A. What barriers could impede access to mental health services for children and adolescents in Africa? A scoping review. BMC Health Serv Res 2023;23(1):348 [FREE Full text] [doi: 10.1186/s12913-023-09294-x] [Medline: 37024835]
14. Kantor V, Knefel M, Lueger-Schuster B. Perceived barriers and facilitators of mental health service utilization in adult trauma survivors: a systematic review. Clin Psychol Rev 2017;52:52-68 [FREE Full text] [doi: 10.1016/j.cpr.2016.12.001] [Medline: 28013081]
15. World Bank Group. Democratic Republic of Congo systematic country diagnostic: policy priorities for poverty reduction and shared prosperity in a post-conflict country and fragile state. Open Knowledge Repository. 2018. URL: https://openknowledge.worldbank.org/entities/publication/c806f0c7-027c-585d-8d48-49241ef863e2 [accessed 2026-01-15]
16. Democratic Republic of Congo. The World Bank in DRC. 2023. URL: https://www.worldbank.org/en/country/drc/overview [accessed 2026-01-06]
17. Kala Kamdjoug JR. Change management and digital transformation project success in SMEs located in the Democratic Republic of the Congo. JEIM 2024;37(2):580-605 [FREE Full text] [doi: 10.1108/jeim-09-2022-0340]
18. Plackett R, Steward J, Kassianos AP, Duenger M, Schartau P, Sheringham J, et al. The effectiveness of social media campaigns in improving knowledge and attitudes toward mental health and help-seeking in high-income countries: scoping review. J Med Internet Res 2025;27:e68124 [FREE Full text] [doi: 10.2196/68124] [Medline: 40408767]
19. Schröder R, Hamer T, Kruzewitz V, Busch E, Suhr R, König L. Effects of YouTube health videos on mental health literacy in adolescents and teachers: randomized controlled trial. JMIR Ment Health 2025;12:e76004 [FREE Full text] [doi: 10.2196/76004] [Medline: 40743523]
20. De Vecchi N, Kenny A, Dickson-Swift V, Kidd S. How digital storytelling is used in mental health: a scoping review. Int J Ment Health Nurs 2016;25(3):183-193. [doi: 10.1111/inm.12206] [Medline: 26900000]
21. Park E, Forhan M, Jones CA. The use of digital storytelling of patients' stories as an approach to translating knowledge: a scoping review. Res Involv Engagem 2021;7(1):58 [FREE Full text] [doi: 10.1186/s40900-021-00305-x] [Medline: 34454604]
22. Lambert C, Egan R, Turner S, Milton M, Khalu M, Lobo R, et al. The digital bytes project: digital storytelling as a tool for challenging stigma and making connections in a forensic mental health setting. Int J Environ Res Public Health 2023;20(13):6268 [FREE Full text] [doi: 10.3390/ijerph20136268] [Medline: 37444115]
23. Brennan S, Geary U, Gallagher S. Online videos promote brain health literacy. Health Promot Int 2021;36(5):1243-1252. [doi: 10.1093/heapro/daaa142] [Medline: 33383581]
24. Xuan W, Tian K, Hao L. Quality assessment of short videos on health science popularization in China: scale development and validation. Front Public Health 2025;13:1640105 [FREE Full text] [doi: 10.3389/fpubh.2025.1640105] [Medline: 41048298]

25. Xiao L, Min H, Wu Y, Zhang J, Ning Y, Long L, et al. Public's preferences for health science popularization short videos in China: a discrete choice experiment. Front Public Health 2023;11:1160629 [FREE Full text] [doi: 10.3389/fpubh.2023.1160629] [Medline: 37601206]
26. Krumm IR, Miles MC, Clay A, Carlos Ii WG, Adamson R. Making effective educational videos for clinical teaching. Chest 2022;161(3):764-772 [FREE Full text] [doi: 10.1016/j.chest.2021.09.015] [Medline: 34587482]
27. Srinivasa K, Charlton A, Moir F, Goodyear-Smith F. How to develop an online video for teaching health procedural skills: tutorial for health educators new to video production. JMIR Med Educ 2024;10:e51740 [FREE Full text] [doi: 10.2196/51740] [Medline: 39110488]
28. Puspitasari IM, Garnisa IT, Sinuraya RK, Witriani W. Perceptions, knowledge, and attitude toward mental health disorders and their treatment among students in an Indonesian University. Psychol Res Behav Manag 2020;13:845-854 [FREE Full text] [doi: 10.2147/PRBM.S274337] [Medline: 33149708]
29. King S, Chen KLD, Chokshi B. Becoming trauma informed: validating a tool to assess health professional's knowledge, attitude, and practice. Pediatr Qual Saf 2019;4(5):e215 [FREE Full text] [doi: 10.1097/pq9.0000000000000215] [Medline: 31745518]
30. Hearn M, Ceschi G, Brillon P, Fürst G, Van der Linden M. A French adaptation of the Posttraumatic Diagnostic Scale. Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement 2012;44(1):16-28. [doi: 10.1037/a0025591]
31. Ertl V, Pfeiffer A, Saile R, Schauer E, Elbert T, Neuner F. Validation of a mental health assessment in an African conflict population. Psychol Assess 2010;22(2):318-324. [doi: 10.1037/a0018810] [Medline: 20528059]
32. McCarthy S. Post-traumatic Stress Diagnostic Scale (PDS). Occup Med (Lond) 2008;58(5):379. [doi: 10.1093/occmed/kqn062] [Medline: 18676430]
33. Kang Y, Mesquiti S, Baik ES, Falk EB. Empathy and helping: the role of affect in response to others' suffering. Sci Rep 2025;15(1):3256 [FREE Full text] [doi: 10.1038/s41598-025-87221-2] [Medline: 39863716]
34. Maffei A, Spironelli C, Angrilli A. Affective and cortical EEG gamma responses to emotional movies in women with high vs low traits of empathy. Neuropsychologia 2019;133:107175. [doi: 10.1016/j.neuropsychologia.2019.107175] [Medline: 31449821]
35. Elam T, Efthemiou A, Taku K. The association positive and negative empathy have with depressive symptoms, resilience, and posttraumatic growth. Sci Rep 2025;15(1):9464 [FREE Full text] [doi: 10.1038/s41598-025-86285-4] [Medline: 40108175]
36. Górska G, Berkovich-Ohana A, Klimecki O, Trautwein F. Situational assessment of empathy and compassion: predicting prosociality using a video-based task. PLoS One 2023;18(12):e0289465 [FREE Full text] [doi: 10.1371/journal.pone.0289465] [Medline: 38060491]
37. Shu J, Hassell S, Weber J, Ochsner KN, Mobbs D. The role of empathy in experiencing vicarious anxiety. J Exp Psychol Gen 2017;146(8):1164-1188. [doi: 10.1037/xge0000335] [Medline: 28627907]
38. Cosme D, Scholz C, Chan H, Doré BP, Pandey P, Carreras-Tartak J, et al. Message self and social relevance increases intentions to share content: correlational and causal evidence from six studies. J Exp Psychol Gen 2023;152(1):253-267. [doi: 10.1037/xge0001270] [Medline: 35951378]
39. Scholz CC, C Baek E, Falk EB. Invoking self-related and social thoughts impacts online information sharing. Soc Cogn Affect Neurosci 2023;18(1):nsad013 [FREE Full text] [doi: 10.1093/scan/nsad013] [Medline: 36869716]
40. Nygård T, Lindfors P. Promoting youth well-being: a qualitative study of Finnish YouTubers' mental health content. Health Promot Int 2025;40(3):daaf074. [doi: 10.1093/heapro/daaf074] [Medline: 40500241]
41. Keselman A, Arnott Smith C, Leroy G, Kaufman DR. Factors influencing willingness to share health misinformation videos on the internet: web-based survey. J Med Internet Res 2021;23(12):e30323 [FREE Full text] [doi: 10.2196/30323] [Medline: 34889750]
42. Cai Y, Kamarudin S, Nujaimi S. Willingness to share information on social media: a systematic literature review (2020-2024). Front Psychol 2025;16:1567506 [FREE Full text] [doi: 10.3389/fpsyg.2025.1567506] [Medline: 40535178]
43. Hu J, Tang K, Qian X, Sun F, Zhou W. Behavioral change in waste separation at source in an international community: an application of the theory of planned behavior. Waste Manag 2021;135:397-408. [doi: 10.1016/j.wasman.2021.09.028] [Medline: 34614467]
44. Kim R, Lin T, Pang G, Liu Y, Tungate AS, Hendry PL, et al. Derivation and validation of risk prediction for posttraumatic stress symptoms following trauma exposure. Psychol Med 2023;53(11):4952-4961. [doi: 10.1017/S003329172200191X] [Medline: 35775366]
45. Ozer EJ, Best SR, Lipsey TL, Weiss DS. Predictors of posttraumatic stress disorder and symptoms in adults: a meta-analysis. Psychol Bull 2003;129(1):52-73. [doi: 10.1037/0033-2909.129.1.52] [Medline: 12555794]
46. Ziobrowski HN, Kennedy CJ, Ustun B, House SL, Beaudoin FL, An X, AURORA Consortium, et al. Development and validation of a model to predict posttraumatic stress disorder and major depression after a motor vehicle collision. JAMA Psychiatry 2021;78(11):1228-1237 [FREE Full text] [doi: 10.1001/jamapsychiatry.2021.2427] [Medline: 34468741]

47. Shaygan M, Yazdani Z, Valibeygi A. The effect of online multimedia psychoeducational interventions on the resilience and perceived stress of hospitalized patients with COVID-19: a pilot cluster randomized parallel-controlled trial. BMC Psychiatry 2021;21(1):93 [FREE Full text] [doi: 10.1186/s12888-021-03085-6] [Medline: 33573631]
48. Rowe H, Sperlich M, Cameron H, Seng J. A Quasi-experimental outcomes analysis of a psychoeducation intervention for pregnant women with abuse-related posttraumatic stress. J Obstet Gynecol Neonatal Nurs 2014;43(3):282-293 [FREE Full text] [doi: 10.1111/1552-6909.12312] [Medline: 24754455]
49. Byrne M, Campos C, Daly S, Lok B, Miles A. The current state of empathy, compassion and person-centred communication training in healthcare: an umbrella review. Patient Educ Couns 2024;119:108063 [FREE Full text] [doi: 10.1016/j.pec.2023.108063] [Medline: 38008647]
50. Moudatsou M, Stavropoulou A, Philalithis A, Koukouli S. The role of empathy in health and social care professionals. Healthcare (Basel) 2020;8(1):26 [FREE Full text] [doi: 10.3390/healthcare8010026] [Medline: 32019104]
51. Hudon A, Perry K, Plate A, Doucet A, Ducharme L, Djona O, et al. Navigating the maze of social media disinformation on psychiatric illness and charting paths to reliable information for mental health professionals: observational study of TikTok videos. J Med Internet Res 2025;27:e64225 [FREE Full text] [doi: 10.2196/64225] [Medline: 40532184]
52. Nicholas A, Joshua O, Elizabeth O. Accessing mental health services in Africa: current state, efforts, challenges and recommendation. Ann Med Surg (Lond) 2022;81:104421 [FREE Full text] [doi: 10.1016/j.amsu.2022.104421] [Medline: 35996570]
53. Sankoh O, Sevalie S, Weston M. Mental health in Africa. Lancet Glob Health 2018;6(9):e954-e955 [FREE Full text] [doi: 10.1016/S2214-109X(18)30303-6] [Medline: 30103990]
54. Munir MM, Ahmed N. Using social media platforms to raise health awareness and increase health education in Pakistan: structural equation modeling analysis and questionnaire study. JMIR Hum Factors 2025;12:e65745 [FREE Full text] [doi: 10.2196/65745] [Medline: 40194316]
55. Alemu WG, Due C, Muir-Cochrane E, Mwanri L, Ziersch A. Internalised stigma among people with mental illness in Africa, pooled effect estimates and subgroup analysis on each domain: systematic review and meta-analysis. BMC Psychiatry 2023;23(1):480 [FREE Full text] [doi: 10.1186/s12888-023-04950-2] [Medline: 37386417]
56. Faleti DD, Akinlotan O. Stigmatisation of mental illness in Africa: a systematic review of qualitative and mixed studies. J Ment Health 2025;34(6):716-733. [doi: 10.1080/09638237.2024.2426982] [Medline: 39576718]
57. Jewkes R, Mhlongo S, Chirwa E, Seedat S, Myers B, Peer N, et al. Pathways to and factors associated with rape stigma experienced by rape survivors in South Africa: analysis of baseline data from a rape cohort. Clin Psychol Psychother 2022;29(1):328-338 [FREE Full text] [doi: 10.1002/cpp.2637] [Medline: 34170058]
58. Willan S, Shai N, Majola T, Mabhida M, Mngadi S, Gounden T, et al. South African rape survivors’ expressions of shame, self-blame and internalized-stigma. SSM - Mental Health 2024;5:100310. [doi: 10.1016/j.ssmmh.2024.100310]
59. Purnell L, Chiu K, Bhutani GE, Grey N, El-Leithy S, Meiser-Stedman R. Clinicians' perspectives on retraumatisation during trauma-focused interventions for post-traumatic stress disorder: a survey of UK mental health professionals. J Anxiety Disord 2024;106:102913 [FREE Full text] [doi: 10.1016/j.janxdis.2024.102913] [Medline: 39111232]
60. Schock K, Rosner R, Wenk-Ansohn M, Knaevelsrud C. Retraumatization--a conceptional approach. Psychother Psychosom Med Psychol 2010;60(7):243-249. [doi: 10.1055/s-0030-1248268] [Medline: 20301050]
61. Trautmann S, Reineboth M, Trikojat K, Richter J, Hagenaars MA, Kanske P, et al. Susceptibility to others' emotions moderates immediate self-reported and biological stress responses to witnessing trauma. Behav Res Ther 2018;110:55-63. [doi: 10.1016/j.brat.2018.09.001] [Medline: 30243101]
62. Huang C, Wu Z, Sha S, Liu C, Yang L, Jiang P, et al. The dark side of empathy: the role of excessive affective empathy in mental health disorders. Biol Psychiatry 2025;98(5):404-415. [doi: 10.1016/j.biopsych.2024.12.020] [Medline: 39793690]
63. Zenzmaier C, Janssen J, Zulmin C, Österreicher P, Heinrich L, Tucek G, et al. Response of salivary biomarkers to an empathy triggering film sequence-a pilot study. Sci Rep 2021;11(1):15794 [FREE Full text] [doi: 10.1038/s41598-021-95337-4] [Medline: 34349165]
64. Schincariol A, Orrù G, Otgaar H, Sartori G, Scarpazza C. Posttraumatic stress disorder (PTSD) prevalence: an umbrella review. Psychol Med 2024;54(15):1-14. [doi: 10.1017/S0033291724002319] [Medline: 39324396]
65. Diamond PR, Airdrie JN, Hiller R, Fraser A, Hiscox LV, Hamilton-Giachritsis C, et al. Change in prevalence of post-traumatic stress disorder in the two years following trauma: a meta-analytic study. Eur J Psychotraumatol 2022;13(1):2066456 [FREE Full text] [doi: 10.1080/20008198.2022.2066456] [Medline: 35646293]
66. Koenen KC, Ratanatharathorn A, Ng L, McLaughlin KA, Bromet EJ, Stein DJ, et al. Posttraumatic stress disorder in the World Mental Health Surveys. Psychol Med 2017;47(13):2260-2274 [FREE Full text] [doi: 10.1017/S0033291717000708] [Medline: 28385165]
67. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991;50(2):179-211 [FREE Full text] [doi: 10.1016/0749-5978(91)90020-T]

Abbreviations

PDS-F: Post-traumatic Diagnostic Scale – French adaptation

PTSD: posttraumatic stress disorder
VIF: variance inflation factor

Edited by J Torous; submitted 25.Jul.2025; peer-reviewed by L Ou, N Ean; comments to author 01.Nov.2025; revised version received 25.Nov.2025; accepted 27.Nov.2025; published 26.Jan.2026.

Please cite as:
Bapolisi A, Foucart J, Kabambi D, Mirishe R, Musa E, Ruvunangiza A, Bosomi J, Bulabula V, Ilunga M, Kajibwami E, Bapolisi O, Nyalundja AD, Igega MH, Mwene-batu P, de Timary P, Khazaal Y
Evaluating a Culturally Tailored Digital Storytelling Intervention to Improve Trauma Awareness in Conflict-Affected Eastern Congo: Quasi-Experimental Pilot Study
JMIR Ment Health 2026;13:e81291
URL: https://mental.jmir.org/2026/1/e81291
doi:10.2196/81291
PMID:

General Practitioners' Perspectives on Digital Health Applications for Mental Disorders and Their Prescribing Behavior: Mixed Methods Study

Sandy Scheibe[1], MPH; Sandra Salm[2,3], PhD; Karola Mergenthal[2], PhD; Deborah Engesser[4], MSc; Esther Stalujanis[4,5,6], PhD; Susanne Singer[4,7], PhD; Pascal Kemmerer[4], MSc; Lena Dotzauer[4], PhD; Karen Voigt[1], PhD

[1]Department of General Practice, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Fetscherstraße 74, 01307 Dresden, Germany

[2]Institute of General Practice, Goethe University Frankfurt, Frankfurt am Main, Germany

[3]Medical Psychology—Neuropsychology and Gender Studies and Center for Neuropsychological Diagnostics and Intervention (CeNDI), Faculty of Medicine and University Hospital Cologne, University of Cologne, Köln, Germany

[4]Division of Epidemiology and Health Services Research, Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

[5]Department of Consultation-Liaison-Psychiatry and Psychosomatic Medicine, University Hospital of Zürich, Zürich, Switzerland

[6]Department of Psychology, Clinical Psychology and Psychotherapy–Methods and Approaches, Trier University, Trier, Germany

[7]Department of Quality of Life in Oncology, Comprehensive Cancer Center Mecklenburg-Vorpommern (CCC-MV), University Medical Centre Rostock, Rostock, Germany

Corresponding Author:
Sandy Scheibe, MPH
Department of General Practice, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Fetscherstraße 74, 01307 Dresden, Germany

Abstract

Background: The high number of mental disorders poses challenges for health care systems. In 2020, digital health applications (DHAs) were introduced in Germany as a new form of health care financed by the statutory health insurance. They aim to detect, monitor, treat, or alleviate disease, injury, or disability. DHAs for mental disorders (DHA-MD) intend to improve outpatient care for patients with mental disorders. However, evidence on general practitioners' (GPs') perspectives on DHA-MD and their prescribing behavior is limited.

Objective: This study aimed to analyze GPs' perspectives on DHA-MD and their prescribing behavior in the care of patients with mental disorders.

Methods: A mixed methods study was conducted (January-October 2024), including a Germany-wide online survey and qualitative interviews with GPs and medical assistants (MAs). Sampling was conducted in collaboration with German research practice networks, which distributed the study invitation to their affiliated GPs. The questionnaire as well as the interview guides for GPs and MAs was developed by the study team according to the Consolidated Framework for Implementation Research. Descriptive analyses of prescribing behavior and perceived need (measured on an 11-point scale) for DHA-MD were conducted, followed by multivariate regression analyses to identify predictors of prescribing behavior and perceived need for DHA-MD. The interviews with GPs and MAs were analyzed using qualitative content analysis according to Mayring.

Results: A sample of 149 GPs participated, and 12 GPs as well as 5 MAs were interviewed. The median prescription frequency of DHA-MD per quarter was 1, whereas the median estimated need was 3. Working in a half digitized and half paper-based practice (odds ratio 5.133, 95% CI 1.695 - 15.542) as well as working in a completely digitized practice (odds ratio 3.006, 95% CI 1.296 - 6.969) positively predicted the prescribing behavior. The duration of GPs' medical practice (b=−0.057; P=.01) negatively predicted the perceived need, while working in a group practice (b=0.980; P=.02) positively predicted the perceived need for DHA-MD. In the interviews, GPs and MAs reported that they valued DHA-MD as a temporary or supplementary option for bridging waiting times for psychotherapy and considered their effectiveness to be highly dependent on indication and patient adherence. Reported barriers of GPs according to DHA-MD included lacking knowledge about DHA-MD, missing effectiveness studies, and difficulties integrating them into existing care processes.

Conclusions: GPs are reluctant to prescribe DHA-MD, as the need is considered to be low and their use is primarily seen as a temporary or supplementary treatment option rather than a stand-alone intervention. There are significant reasons for rejection and barriers that hinder prescription in primary care. Addressing these barriers and involving GPs as well as patients in future research are essential for the development of DHA-MD.

(JMIR Ment Health 2026;13:e78659) doi:10.2196/78659

KEYWORDS

digital health; mobile apps; mental health; mixed methods; primary care; general practitioners

Introduction

The global burden of mental disorders is high, affecting about 970.1 million people in 2019. In Germany, the lifetime prevalence of mental disorders in adults is 25.2% [1], which corresponds to a value in the midrange of international estimates. Mental disorders are among the disease groups with the highest burden for those affected [2,3], including reduced quality of life and functional capacity, inability to work, or early retirement, as well as an increased mortality [4-6]. In 2023, mental disorders accounted for 16.0% of incapacity to work days in Germany, and they were the leading cause of early retirement, accounting for 41.8% [7,8]. In addition, the indirect and direct costs arising from mental disorders are estimated to amount to more than €600 billion (approximately equal to US $696 billion) per year, exceeding about 4.0% of the gross domestic product across the European countries [9]. In Germany, the direct health care costs for mental disorders amounted to around €56.3 million (approximately equal to US $65.3 million) in 2020, showing a significant increase compared with 2015 (€42.7 million, approximately equal to US $49.5 million).

These findings highlight the need for adequate and accessible treatment. However, globally, the treatment coverage for mental disorders is low, suggesting a considerable degree of unmet treatment need [10,11]. Among treated cases with mental disorders, there is a long delay between onset and first treatment contact [10-13]—median waiting times for access to psychotherapy in European countries are more than 2 months [14]. In Germany, the recent psychotherapist act has potentially influenced the waiting times for the first appointment after initial consultation— being 2.6 before and 3.8 months after the reform [15].

To counteract the strained treatment situation, great hopes were being placed in digital health, such as mobile health (mHealth) apps. In 2020, Germany became the first country worldwide to grant statutorily insured individuals the right to use certain mHealth apps at the expense of health insurers. Afterward, other European countries such as France and Belgium have introduced approval procedures for digital health application (DHA)–equivalent health apps similar to the German model [16,17].

DHA are certified medical products, according to the European Medical Device Regulation, primarily based on digital functions to detect, monitor, treat, or alleviate disease, injury, or disability. For legal authorization as a medical device, DHA must demonstrate a positive health care effect (ie, improvement of disease symptoms, quality of life, health literacy, and feeling better involved in the treatment) to be permanently approved by the German Federal Institute for Drugs and Medical Devices (German, Bundesinstitut für Arzneimittel und Medizinprodukte [BfArM]) [18]. To date (October 2025), 56 DHAs are reimbursable by statutory health insurers. Mental disorders represent the largest group of indications for available DHA (29/56, 51.8%), including applications for treatment of insomnia, depression, anxiety disorders, or nicotine dependence (hereinafter referred to as DHA for mental disorders, DHA-MD) [19].

Health care professionals in the outpatient care sector play an important role in the implementation process, as reimbursement of DHA is possible when prescribed by physicians or psychotherapists [20-22]. Four years after the introduction of DHA, they were prescribed by 12.0% of physicians and psychotherapists [23]. The majority of DHA prescriptions were issued by general practitioners (GPs) [24].

Systematic reviews including studies from all over the world analyzing the implementation of digital health technologies in routine care identified several barriers faced by health care professionals, including technical limitations (eg, insufficient network coverage and lack of existing technologies or devices), insufficient expertise, and legal and ethical concerns (eg, privacy and security concerns and national legislation), as well as financial barriers (eg, high costs and inadequate remuneration structure). Reported facilitators included access to reliable information about digital health services, the perceived usefulness, and government monetary incentives [25-27]. However, studies that examine perspectives of health care professionals toward DHA, according to the definition of BfArM, are rare. Dalhausen and colleagues [20] conducted a mixed methods study to examine attitudes of GPs and psychotherapists toward DHA. The results indicated that GPs and psychotherapists expressed a generally positive attitude and openness toward DHAs. Attitudes and prescription intentions were significantly influenced by digital affinity, that is, GPs with a higher digital affinity were more likely to prescribe DHA. Age, practice type, and practice location were not associated with DHA prescriptions.

Another perspective is provided by Posselt and colleagues [21], who examined GPs' key challenges in prescribing DHA-MD for patients with depression. They identified the following challenges: information gaps, insufficient knowledge about available information sources for DHA-MD, and difficulties in selecting patients suitable for DHA-MD use [21].

Previous studies on DHA-MD in Germany predominantly relied on qualitative data or focused on specific indications, such as depression. To date, no study combined quantitative and qualitative research methods to comprehensively analyze GPs' perspectives on DHA-MD, across the whole spectrum of mental disorders. The aim of this study was therefore to analyze GPs' perspectives on DHA-MD and their prescribing behavior in the care of patients with mental disorders.

Methods

Study Design

This study was conducted in Germany. We used a mixed methods convergent parallel design in which quantitative and qualitative data were collected and analyzed separately. The triangulation of the results took place during the interpretation phase by comparing and contrasting the findings of both strands to identify areas of convergence, divergence, or complementarity [28,29].

In the "Results" section, both data strands were first presented separately according to their respective methods, while in the "Discussion" section, the results were integrated and interpreted jointly along overarching themes [30], with qualitative results used to explain and deepen the understanding of quantitative results on GPs' perspectives and prescribing behavior.

Online Survey

Participants and Recruitment

Participants were practicing GPs (specialists in family or internal medicine) working in their own practices or as an employee in a practice, as well as clinical residents. The link to the online survey was distributed nationwide via email to German research practice networks and professional associations, as well as presented at public events for GPs in Germany. Subsequently, the research practice networks forwarded the invitation to their affiliated GPs, resembling a snowball sampling approach. Due to this strategy, it is not possible to determine the exact number of GPs who received the invitation to participate. The data were collected cross-sectionally through an anonymous online survey from January to October 2024.

Questionnaire

The questionnaire was developed by the study team according to the Consolidated Framework for Implementation Research model, which captures barriers and enablers in the implementation of interventions [31]. We conducted a pretest of the questionnaire with 6 GPs who verified the relevance and completeness of the items. Based on the feedback of GPs, the wording of 1 question was revised to improve clarity. The overall structure and methodology of the questionnaire were retained after pretest.

Demographic and Practice Characteristics

We collected sociodemographic information (ie, gender, age, and duration of practicing as GP) and characteristics of GP practices (ie, type of practice [single vs group practice], location [urban-rural], number of treated patients per quarter, and extent of digitization in practice). To determine the extent of digitization in practice, GPs stated on a 5-point scale whether their communication with colleagues in outpatient care and patients is "almost completely digitized" (0) or "almost completely paper-based" (4). For regression analyses, we recoded this variable into 3 categories: "digitized" (0 - 1), "half digitized and half paper-based" (2), and "paper-based" (3-4).

Prescription Behavior and Perceived Need for DHA-MD

Participants were asked how many patients they had prescribed a DHA (all indications) or a DHA-MD in the last quarter. Since a large proportion of participants reported having prescribed no DHA-MD at all, we recoded this variable into a binary measure (prescription=yes or no) to facilitate analysis using logistic regression. A prescription was considered present if at least 1 patient had received a DHA-MD during the last 3 months. The perceived need for DHA-MD for patients was estimated on an 11-point scale, ranging from "very low" (0) to "very high" (10).

Statistical Analysis of the Survey Data

Quantitative data were analyzed using SPSS (version 30; IBM Corp). Descriptive statistics were presented as mean values with standard deviations for metric-scaled variables and as percentages and frequencies for nonmetric-scaled variables in order to describe the study sample.

The following variables were the predictors of the regression analyses: gender (male or female), age (years), duration of GP practice (years), type of practice (single practice or joint practice), practice location (city or rural area), and degree of digitization in practice (paper-based, or half digitized and half paper-based, or digitized). Correlations between predictor variables were examined to rule out multicollinearity. In cases of high correlation (r>0.80) [32], a decision was made to exclude one of the variables from the analysis based on theoretical considerations and relevance to the research question. Dependent variables were analyzed via multivariate analyses using logistic regression modeling for prescribing behavior and linear regression modeling for perceived need of DHA-MD. Missing values were not imputed. Participants with missing data were excluded from the respective analyses. The Gauss-Markov assumptions were tested as prerequisites for the multiple linear regression of perceived need of DHA-MD. Model fit was assessed using Nagelkerke's R^2 for the logistic regression model and adjusted R^2 for the multiple linear regression model. The significance level was set at $\alpha \leq .05$.

Qualitative Study

Participants and Procedures

Invitations for the telephone interviews were distributed via the German research practice network "SaxoForN" [33], which subsequently forwarded the invitation to their affiliated GPs (snowball sampling). Medical assistants (MAs) were also invited to participate in qualitative interviews, as they are closely involved in administrative processes of GP practices and can provide organizational and time relief for GPs in the German health care system [34]. Previous research has shown that patients have a high level of trust in MAs [35], who often serve as the first point of contact for questions or issues that may not be raised during the GP consultation—possibly also with regard to the prescription and use of DHA. Moreover, as certain tasks in primary care are delegated from GPs to MAs [35], we aimed to explore which concrete tasks they take over in relation to DHA—such as assisting in administrative procedures, supporting patient onboarding, or addressing patient inquiries during the usage phase. Including MAs in the qualitative study enabled us to explore their role in the implementation process and capture patient-related challenges they identified, which may extend beyond the GPs" perspective. Only MAs who worked in GP practices with DHA prescriptions were interviewed to ensure whether they were already familiar with the concept of DHA. Interviews were conducted with all interested GPs and MAs who had registered to participate. Telephone interviews took place between July and October 2024 and were conducted by 2 researchers (SaS and SSch) of the study team. Both interviewers were female and health scientists. They had no prior relationship with the participants.

Telephone Interviews

Semistructured interview guides for GPs and MAs were developed according to the Consolidated Framework for Implementation Research framework [31]. The interview guides were pretested by 2 GPs in December 2023. Feedback revealed that no changes were deemed necessary in terms of methodology, structure, or questions.

The interviews with GPs started with an opening question, which transitioned to the following 3 key topics: experiences with DHA (overall; for DHA-MD), attitudes toward DHA-MD, and implementation factors and conditions for the use of DHA-MD in outpatient care. MAs were asked about their tasks related to DHA in GP practice, frequently asked questions of patients related to DHA, and feedback of patients to DHA use.

If necessary, the interviewers asked further questions to go more in-depth on the information the participants provided. Despite the interview guide, participants were encouraged to talk freely without too much interruption from the interviewers. Theoretical saturation was assessed iteratively during data collection by the interviewers and was deemed achieved when interviews no longer yielded novel insights and a sufficient heterogeneity of perspectives had been captured. To ensure a transparent and consensual process, regular team meetings were held to discuss emerging themes and determine the point of saturation. The telephone interviews were audio-recorded and transcribed verbatim. To ensure data protection, all personal data were pseudonymized in the transcription process.

Qualitative Content Analysis of the Interviews

The interviews with GPs and MAs were analyzed using qualitative content analysis according to Mayring [36]. The coding schemes for the interviews were developed using a deductive-inductive approach, based on the previously developed interview guides. Two researchers (SSch and SaS) analyzed the interviews independently with use of the software MAXQDA (version 2020; VERBI Software). The results were subsequently cross-compared, whereby disagreement was discussed until consensus was reached. If necessary, a third senior researcher was consulted.

Ethical Considerations

This study was approved by the ethics committee of State Medical Association of Rhineland-Palatinate (no. 2023‐17268), ethics committee of Dresden University of Technology (no. SR-EK-418092023), and the Medical Faculty of Goethe University Frankfurt am Main (no. 2023‐1505). Participation in the online survey was anonymous and therefore required no consent for the use of data according to German law. Before conducting the telephone interviews, written informed consent was obtained from all participants. All interview data were collected in pseudonymized form. Any identifying information mentioned during the interviews (eg, names of persons, places, or organizations) was anonymized during transcription and replaced with nonidentifiable character strings to prevent any possibility of reidentification. Participants received a compensation of €50 (approximately equal to US $58) for participating in the interview.

Results

Online Survey

Characteristics of the Study Population

A total of 149 participants completed the questionnaire, of whom 47.7% (71/149) were male and 52.3% (78/149) were female. As shown in Table 1, the mean age of the respondents was 50.7 (SD 10.5) years. The mean work experience of GP was 15.2 (SD 9.9) years. Most respondents were practice owners (111/149, 74.5%) and located in urban areas (122/148, 82.4%) with various community sizes (Table 1). About half of the respondents were active in single practices without physician colleagues (74/145, 51.1%), while the other half (71/145, 48.9%) worked jointly with at least 1 colleague.

Table . Description of study population.

Variables	Participants
Categorical variables, n (%)	
Sex, n=148	
Male	71 (47.7)
Female	78 (52.3)
Position in practice, n=149	
Practice owner	111 (74.5)
Employed	31 (20.8)
Clinical residents	7 (4.7)
Type of practice, n=145	
Single practice	74 (51.1)
Joint practice	71 (48.9)
Practice location, n=148	
Large city (>100,000 inhabitants)	50 (33.8)
Medium-sized city (20,000 - 100,000 inhabitants)	35 (23.6)
Small city (5000 - 20,000 inhabitants)	37 (25.0)
Rural community	26 (17.6)
Communication with patients and colleagues in practice, n=141	
Digitized	26 (18.4)
Half digitized and half paper-based	62 (44.0)
Paper-based	53 (37.6)
Treated patients per quarter, n=145	
<1000	18 (12.2)
1000 - 1999	75 (50.6)
2000 - 2999	33 (22.3)
>3000	19 (12.9)
Numerical variables (mean SD)	
Age (years), n=149	50.7 (10.5)
Duration of GP[a] practice (years), n=148	15.2 (10.0)

[a]GP: general practitioner.

Prescribing Behavior and Perceived Need for DHA-MD

Of the participating GPs, 65.7% (90/137) prescribed at least 1 DHA in the last quarter. The median prescription frequency for DHA (for all indications) per quarter was 2 (IQR 0 - 5) and for DHA-MD, it was 1 (IQR 0 - 2). Nearly half of the respondents (68/137, 49.6%) did not prescribe any DHA-MD in the last quarter. The median estimated need was 3 (IQR 1 - 5).

The multivariate logistic regression included the predictors gender, age, duration of GP practice, type of practice, practice location, and degree of digitization in practice. Correlations between predictor variables indicated a high correlation between age and duration of GP practice (r=0.87; P<.001), justifying the decision to exclude age as a predictor in the regression model. All other correlations were low (r<.25). Working in a half digitized and half paper-based practice (odds ratio 5.133, 95% CI 1.695 - 15.542) as well as working in an almost completely digitized practice (odds ratio 3.006, 95% CI 1.296 - 6.969) positively predicted prescribing behavior (Table 2). Hosmer-Lemeshow test indicated a good model fit (χ^2_8=12.43; P>.05) for the logistic regression model. The model explained 13.9% of the variance (Nagelkerke's R^2=0.139).

The multivariate linear regression analysis included the same predictors as in the logistic regression analysis. The Breusch-Pagan test indicated no heteroscedasticity (P=.45). The duration of GPs' medical practice (b=−0.057; P=.01) negatively predicted the perceived need, while working in a group practice (b=0.980; P=.02) positively predicted the perceived need for DHA-MD (Table 3). The model explained 8.2% of the variance (corrected R^2=0.082; P=.009).

Table . Multivariate logistic regression analysis for general practitioners' prescribing behavior (yes/no) of digital health applications for people with mental disorders (n=133).

Predictor	OR[a] (95% CI)	*P* value
Sex		
Female	Reference category	
Male	0.608 (0.285 - 1.296)	.20
Duration of GP[b] practice	0.992 (0.953 - 1.032)	.69
Type of practice		
Single practice	Reference category	
Joint practice	1.788 (0.835 - 3.830)	.14
Practice location		
City	Reference category	
Rural area	1.177 (0.556 - 2.490)	.67
Degree of digitization in practice		
Paper-based	Reference category	
Half paper-based/half digitized	**5.133** [c] **(1.695 - 15.542)**	**.004**
Digitized	**3.006 (1.296 - 6.969)**	**.01**

[a]OR: odds ratio.

[b]GP: general practitioner.

[c]Values in boldface indicate statistical significance.

Table . Multivariate linear regression analysis for general practitioners' perceived need of digital health applications for people with mental disorders (n=135).

Predictor	*b* value	95% CI	*P* value
Sex			
Male	Reference category		
Female	0.350	−0.498 to 1.199	.45
Duration of GP[a] practice	**−0.057** [b]	**−0.102 to −0.012**	**.01**
Type of practice			
Single practice	Reference category		
Joint practice	**0.980**	**0.138 to 1.822**	**.02**
Practice location			
City	Reference category		
Rural area	0.227	−0.629 to 1.082	.60
Degree of digitization in practice			
Paper-based	−0.823	−1.775 to 0.128	.09
Half paper-based/half digitized	Reference category		
Digitized	0.257	−0.888 to 1.402	.66

[a]GP: general practitioner.

[b]Values in boldface indicate statistical significance.

Telephone Interviews

Interviews were conducted with 12 GPs and 5 MAs. The interviews with GPs varied in duration between 18 and 38 minutes (mean duration 25 minutes), and the interviews with MAs varied between 7 and 13 minutes (mean duration 9 minutes).

GPs' Experiences With DHA-MD

All interviewed GPs stated that they had *already prescribed DHA* for different indications in practice, including DHA-MD addressing mental disorders. *Reasons for prescribing* DHA-MD to patients were positive experiences with individual DHA-MD, patient request, the ability for GPs to view the contents before prescribing (GP test access), and the opportunity to offer patients a treatment alternative to existing therapy options.

GPs' e*xperiences with health insurance companies*, as well as with the *activation process of DHA-MD*, were mixed. Some GPs mentioned occasional problems with health insurance companies, including long waiting times for prescription processing, prescriptions being completely rejected by individual health insurance companies, and technical problems when redeeming them. Others stated that the processes with health insurance companies and the activation process of DHA-MD were straightforward, and patients usually gained access to DHA-MD within a few days (Multimedia Appendix 1).

Feedback from GPs regarding their *experiences with the utilization of DHA-MD* by patients was heterogeneous. Some GPs reported positive user experiences of patients. Others, however, felt that only a certain group of patients benefited from DHA and used it as intended —namely, those who were particularly motivated and willing to actively engage in the management of their mental health condition. In a few cases, GPs reported that they issued prescriptions, but these were not redeemed at all by patients. GPs stated that patients carefully consider whether treatment with DHA-MD fits their personal context and preferences and whether the content of the DHA-MD aligns with their needs. In addition, GPs reported a lack of integration in the treatment with the DHA-MD, making it difficult for them to monitor patient adherence and the treatment progress.

GPs' Attitudes to DHA-MD

Regarding the *assessment of the effectiveness of DHA-MD*, 4 different groups emerged in the analysis. The first group of GPs, who considered *DHA-MD to be effective and without side effects*, emphasized that DHA-MD would have stabilizing effects on patients by providing validated knowledge about their mental disorder and is especially helpful in bridging the waiting time for psychotherapy appointments. GPs further stated that DHA-MD could reduce physician-patient contacts and, in some mild cases, make psychotherapeutic treatment no longer necessary. The second group of GPs stated that DHA-MD *effectiveness depends on the patient characteristics*. Patients must be motivated and should use the DHA-MD as prescribed by the manufacturers to achieve sufficient effectiveness. The third group of GPs declared that DHA-MD *effectiveness depends on the indication*. According to the GPs, DHA-MD would be effective for conditions where cognitive-behavioral therapy is effective, such as anxiety disorders, but less so for borderline personality disorders. In addition, GPs mentioned that providing proof of effectiveness might be more difficult for certain mental health conditions because it cannot be measured as objectively as in other conditions (eg, weight loss in obesity). The fourth group of GPs expressed that evidence-based statements about the effectiveness of DHA-MD cannot be made due to *insufficient evidence*. They highlighted the lack of long-term studies and large-scale cohort studies, including subgroup analyses, to assess effectiveness in different patient groups.

With regard to *the importance of DHA-MD* in primary care, GPs mentioned societal benefits, meaning that the treatment with DHA-MD could reduce the use of antidepressants, minimize sick leave, and ease the burden on GP practices by reducing patient-physician contacts.

Regarding the question of how GPs assessed the potential of DHA-MD in the *collaboration between GPs and psychotherapists*, 2 different groups emerged. The first group of GPs emphasized that there has been little exchange between GPs and psychotherapists so far, and they believed that this situation would not change even with the development and integration of these innovative digital technologies into patient care. A second group of GPs could imagine that both groups of health care providers would be included in the DHA-MD, allowing them to simultaneously track the treatment process and promote interdisciplinary collaboration.

GPs' assessment of *the perceived need for DHA-MD in primary care* was heterogeneous. On the one hand, GPs stated that there is a need for DHA-MD, justified particularly due to an increase in mental disorders, especially among younger adults, and as a consequence of the COVID-19 pandemic. In addition, they reported that DHA-MD would be helpful if they addressed a problem in general practice—for example, by providing treatment for patients in need of psychotherapy who face long waiting times, thereby alleviating pressure on GP practices. However, GPs also highlighted that developing new DHA-MD for indications already covered by existing products is unnecessary, as the market is becoming increasingly opaque, and they lack the time in their daily practice to inform themselves comprehensively about new products. Instead, they advocated in the interviews for a reevaluation and regular improvement of the quality of existing DHA-MD. On the other hand, GPs stated that the evidence for practical use of DHA-MD is still insufficient and that these digital applications would not be necessary if there were enough psychotherapy places and enough medical capacity that the conversations with patients could be partly conducted by GPs themselves.

MAs' Experiences With DHA

Overall, MAs' *experiences with DHA* were heterogeneous (Multimedia Appendix 2). All interviewed MAs stated that DHA had *already been prescribed* in the GP practices where they worked, including DHA for musculoskeletal disorders, mental disorders, and metabolic disorders. According to MAs, their main *sources of information about DHA* were the internet, manufacturer advertising, specialist journals, and test access. However, all MAs reported that they *lacked sufficient knowledge* about the contents and functionalities of DHA and expressed a need for increased public relations work and training in order to better assess the functionality of DHA, as well as the suitability for individual patients. MAs also reported *challenges in selecting patients suitable for DHA-MD use* because patients often do not talk openly about their mental disorder due to feelings of shame.

MAs evaluated *the importance of DHA* in primary care differently. On the one hand, DHAs were perceived as an innovative care solution, particularly for bridging waiting times for psychotherapy and as a supportive therapy. On the other hand, doubts were expressed about their usefulness, since personal contact to a GP or a psychotherapist was considered essential for mental disorders. With regard to the *integration of DHA into GP practice*, MAs reported that implementation in rural areas with a predominantly older patient population is difficult because they tend to be less tech-savvy. Most MAs reported not taking on any *tasks related to DHA* in GP practice. In some cases, there were organizational tasks for MAs in relation to DHA, for instance, preparing flyers for interested patients or reminding GPs to use the billing code in the reimbursement process when a DHA was prescribed. Most MAs reported that they had not been asked any *questions by patients in connection with DHA*. There were single questions regarding the functionality of DHA, the necessity to use, and if patients specifically requested a DHA.

According to MAs, detailed questions of patients were addressed directly with the GPs during the consultations. Regarding the *feedback of patients to DHA*, which was shared with MAs, three different groups emerged: (1) positive feedback, (2) negative feedback, and (3) no feedback received. The first group reported positive user experiences. According to MAs, younger patients who were highly motivated to use DHA benefit particularly and receive support in dealing with their own disorder, as well as making sustainable lifestyle changes. The second group reported negative feedback in detail that patients did not get along with the use of DHA. There was a third group of MAs who stated that they did not receive any feedback from patients because this was discussed exclusively during consultations between the GP and the patient.

Discussion

Principal Findings

Our study aimed to analyze GPs' perspectives on DHA-MD and their prescribing behavior in the care of patients with mental disorders. The study is the first in Germany to show that while the majority of participating GPs have already prescribed DHA, they tend to prescribe DHA-MD only selectively in primary care and perceive their need as low. The importance of DHA-MD is particularly seen in bridging waiting times for psychotherapy. According to GPs, there are considerable reasons for rejection as well as barriers that hinder prescription.

Prescribing Behavior

The proportion of GPs who have already prescribed DHA in our study was 65.7%, which was even higher than that in other studies from various countries (7.9% [20], 31.0% [37], and 50.0% [38]). Current billing data from German health insurance companies indicated a steady increase in DHA prescriptions with subsequent patient usage since 2020, rising from 41,000 to 209,000 in 2023 [24]. The relatively low prescription rate of 7.9% in the study by Dalhausen et al [20], conducted in Germany, may be explained by the study period, as the concept of DHA was still relatively new at this time. In contrast, the 50.0% rate reported in an Australian study [38] reflects a different health care context and a broader focus on mHealth app use rather than DHA-specific prescriptions. In addition, in Australia, there is no national reimbursement system for mHealth apps, unlike in Germany, meaning that patients have to privately fund these health applications [38].

Focusing especially on DHA-MD, half of the participating GPs in our study had prescribed at least 1 DHA-MD in the last 3 months. Mental disorders are one of the most common reasons for consultation in general practice [39]. In this context, billing data from German health insurance companies revealed that the majority of DHA-MD prescriptions (45%) were issued by GPs [24]. The interviews with GPs provided a possible explanation, indicating that GPs prescribe DHA-MD particularly to offer their patients an alternative treatment option and to bridge waiting times for psychotherapy.

Furthermore, existing literature highlighted that health care professionals are more likely to adopt a new technology if they perceive it as beneficial for their own work or their patients [22,40]. Consistent with these findings, GPs in our study have not only emphasized patient-related advantages of DHA-MD, such as improved patient education and the low-threshold access, but also pointed to disadvantages including delayed physician-patient interaction and adherence problems, which could explain the nonprescription rate as well as the low median prescription frequency of DHA-MD in our study. A potential selection bias toward increased participation of DHA critical GPs is also conceivable.

The multivariate logistic regression analysis identified the degree of digitization as a significant predictor of prescribing behavior, which aligns with the findings by Dalhausen et al [20] and international systematic reviews [22,40,41] on the adoption of digital health technologies, showing that health care professionals with greater digital affinity and experience held significantly more positive attitudes and were more likely to adopt digital technologies in practice. A possible explanation could be that GPs working in practices with a higher extent of digitization may experience fewer barriers to integration and greater confidence in practical benefits of DHA-MD.

However, both of our regression models showed limited explanatory power, suggesting that additional variables not included in the analyses may also contribute to explain GPs' prescribing behavior and the perceived need. Future studies may therefore include other established predictors of digital health utilization, such as previous training, digital skills, general acceptance of and attitudes toward technology, or knowledge and beliefs about the intervention [42-45], to further elucidate these underlying mechanisms.

GPs' Perspectives on DHA-MD

About two-thirds of the survey respondents assessed the need for DHA-MD as low, which corresponds to a German study by Wangler and Jansky [46] on attitudes and experiences of GPs, indicating that some GPs refrain from prescribing DHA, considering their contribution to the improvement of health minimal. Our interviews revealed reasons for rejection of DHA-MD, including lacking knowledge about available DHA-MD, missing evidence on DHA-MD effectiveness, and limited integrability into existing care processes. These barriers align with a German systematic review on incentives for DHAs among physicians and psychotherapists [27], and an international systematic review on digital health services for musculoskeletal conditions in primary care [26], which additionally identified barriers related to data security and protection, the organizational workload, and the negative impact on the doctor-patient relationship. As in our interviews, high costs, inadequate reimbursement, missing financial incentives, and unclear liability risks were further reported as barriers [27]. A possible solution to these perceived barriers could be the implementation of “digital navigators” [47]—specially trained MAs who support health care professionals by evaluating available DHA, selecting suitable applications for patients, and preparing app-generated data for clinical decision-making. However, evidence on acceptance by health care professionals and feasibility of implementing these digital navigators in outpatient care has not yet been published [47], and further randomized controlled studies are needed to evaluate (cost)-effectiveness.

Our qualitative interviews yielded mixed perceptions of the need for DHA-MD. On the one hand, GPs stated that there would be no need for DHA-MD if sufficient psychotherapy places or adequate medical capacity were available. On the other hand, the majority of interviewed GPs reported a need for DHA-MD, particularly in light of the perceived increase in the number of patients with mental disorders due to the COVID-19 pandemic. There are no published comparable studies available, which quantify the perceived need of DHA-MD from the perspective of GPs, so our results close this evidence gap.

Our study found that perceived need declined with increasing years of GP experience. Due to the high correlation between age and duration of GP practice, age was not included as a predictor in the linear regression analysis. However, as these 2 variables are strongly related both statistically and conceptually, and evidence on the influence of duration of GP practice on the perceived need is lacking, available studies examining age-related effects on attitudes and health care technology adoption may offer valuable context for interpreting our findings. In accordance with our results, available literature showed that, in Germany, younger GPs rate DHA more positively than older GPs [20,48]. This age-related trend is also reflected in international studies from Australia [49] and Brazil [50] and could be explained by the fact that younger physicians are less hesitant to integrate new technologies into their practice [51].

The linear regression analysis revealed that GPs working in group practices rated the need for DHA-MD higher than those in single practices, which is in line with a systematic review from the United States on the adoption and use of health information technology in physician practice organizations. The results showed that compared with single practices, groups with 4-6 physicians were more likely to have an electronic medical record [52,53]. Peer influence and the organizational culture may contribute to this finding. Pollack and colleagues [54] showed in their study that social contagion among physicians has a significant influence on technology adoption—the more closely and frequently physicians interact with colleagues who also use a specific technology, the more likely it is that they will use it themselves. Furthermore, the international systematic review by Police et al [52] highlighted that a lack of commitment to technology integration, along with an organizational culture resistant to change, significantly impedes technology integration and utilization.

While our results indicated that GPs and their MAs preselect patients for the use of DHA-MD, the actual utilization is shaped by patients’ social determinants, which may limit access for certain patient groups and lead to systematic differences in digital participation (def. “digital divide”) [55,56]. According to the World Health Organization’s scoping review, people with greater health care needs, older adults, and marginalized groups are less likely to benefit from digital health interventions. In contrast, younger individuals with higher socioeconomic status living in urban areas tend to experience more positive effects when using these digital technologies [57]. These findings underline the necessity of future effectiveness studies on DHA-MD that incorporate subgroup analyses to assess not only which patient groups benefit most but also which may experience adverse effects or even harms.

To mitigate these disparities and ensure equitable access, it is crucial to integrate both health care providers and patients in the development of DHA-MD through a participatory approach, for instance, Co-Creation, to address their perceived barriers and to ensure that these technologies are designed, developed, and implemented within the specific contexts in which they will be applied [58]. In psychiatric care, this participatory process is particularly valuable to address the complex needs and challenges faced by users in their everyday lives. A participatory development process can significantly increase the acceptance and trust of both users and health care providers, who prescribe DHA-MD to their patients.

Strengths and Limitations

Our study provides information to understand perspectives of GPs on DHA-MD and their prescribing behavior more comprehensively. By using a mixed methods design, we were able to triangulate quantitative data identifying broader trends with qualitative data that provide a more detailed and heuristic understanding of individual GP and MA perspectives. This methodological approach ensured that our findings are relevant for clinical practice, as well as for other health care stakeholders (eg, DHA manufacturer, BfArM). Although conducting an anonymous online survey may have introduced self-selection bias—with participation potentially skewed toward individuals with preexisting interest or strong opinions about DHA-MD—and qualitative interviews primarily attracted already interested participants, the mixed methods design addressed these limitations. The combination of quantitative and qualitative data allowed validation and mutual supplementation of the data. This integration mitigated the biases inherent in the individual methods and strengthened the robustness and practical relevance of the overall findings. However, a limitation of the qualitative study is that we conducted only 5 interviews with MAs. These interviews were intended to understand MAs’ role in the prescription process and use of DHA and to capture perspectives that might extend beyond those of GPs, for example, questions or issues raised by patients to MAs that are not discussed during GP consultations. However, most of the interviewed MAs reported no involvement in DHA prescription or follow-up. Therefore, these findings should be interpreted as exploratory. With regard to generalizability, it is important to acknowledge that our findings are embedded within the specific context of DHA in Germany. The German DHA framework provides a distinct legal foundation for the prescription and reimbursement of such applications. Nonetheless, the results may also hold relevance for other countries that have recently implemented comparable initiatives, including France, Belgium, and the United Kingdom. However, differences between national health care systems and reimbursement structures may constrain the direct transferability of our findings. Consequently, further international research is warranted to explore whether the patterns identified in our study can be replicated in other health care contexts. As the study used a cross-sectional design, it is inherently limited in its ability to infer causal relationships between the variables analyzed, such as between prescription frequency, digital affinity, and age.

Conclusions

DHA-MD are currently prescribed cautiously by GPs, and their perceived need for patients with mental disorders is considered low, as reflected in the relatively high nonprescription rate observed in our study. GPs primarily justify the prescription of DHA-MD as a temporary solution to bridge waiting times for psychotherapy appointments or as a supplementary therapy option, rather than as a stand-alone intervention.

According to GPs, there are reasons for rejection as well as considerable barriers, primarily related to the structural framework of the DHA concept, which hinder prescription of DHA-MD in primary care. Given GPs' key role in the prescription process, addressing both their perceived barriers and those of patients, as end users, is essential for the development of DHA-MD. One possible solution could be to actively involve both patients and health care providers in the development of DHA-MD through a Co-Creation approach to ensure that DHAs are need-related and designed within the specific health care settings in which they are used.

As the digital health care landscape continues to evolve rapidly—driven by technological advancements and shifting health care needs that frequently reshape regulatory frameworks and the availability of DHA-MD—ongoing research on GPs' perspectives on DHA-MD is essential. In particular, future effectiveness studies are needed to objectively evaluate not only which patient groups benefit and which may even be harmed when using DHA-MD but also where alternative therapy approaches (eg, primary or psychotherapy care) are more effective.

Acknowledgments

The authors used the generative artificial intelligence tool ChatGPT by OpenAI for language refinement and drafting suggestions during the revision process. They critically reviewed all outputs and revised them accordingly and stated that they are solely responsible for the content of this paper.

Funding

This study was funded by the Innovation fund of the Federal Joint Committee (project no. 01VSF22022). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: S Scheibe, DE, LD, S Salm, KM, KV, S Singer
Data curation: S Scheibe, S Salm, DE
Formal analysis: S Scheibe
Funding acquisition: S Singer, KM, KV
Investigation: S Scheibe, S Salm
Methodology: S Scheibe, S Salm, KM, KV
Project administration: S Scheibe, S Salm, ES, LD
Resources: S Singer, KM, KV
Supervision: S Singer, KM, KV
Validation: S Scheibe, S Salm, KV
Visualization: S Scheibe
Writing original draft: S Scheibe
Writing–Review & Editing: ES, DE, PK, LD, S Salm, S Scheibe, KM, KV, S Singer

Conflicts of Interest

S Scheibe worked from 2021 to 2023 in a company ("WIG2 institute") that supports manufacturers of digital health applications in the development and authorization process of their applications. S Singer received honoraria for her work as a referee for the

Lilly Quality of Life Award, outside of this study. S Salm received an honorarium from the Austrian Health Insurance Fund for giving a lecture. All other authors do not declare conflicts of interest.

Multimedia Appendix 1
Results of interviews with general practitioners.
[DOCX File, 16 KB - mental_v13i1e78659_app1.docx]

Multimedia Appendix 2
Results of interviews with medical assistants.
[DOCX File, 18 KB - mental_v13i1e78659_app2.docx]

References

1. Kessler RC, Angermeyer M, Anthony JC, et al. Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. World Psychiatry 2007 Oct;6(3):168-176. [Medline: 18188442]
2. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150. [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
3. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. Lancet Psychiatry 2016 Feb;3(2):171-178. [doi: 10.1016/S2215-0366(15)00505-2] [Medline: 26851330]
4. Cohrdes C, Hapke U, Nübel J, Thom J, Erkennen-Bewerten-Handeln. Schwerpunktbericht zur psychischen Gesundheit der Bevölkerung in Deutschland. Teil 1–Erwachsenenalter [Recognize–assess–act: focus report on the mental health of the population in Germany. Part 1–adulthood]. : Robert Koch-Institut; 2022. [doi: 10.25646/8831]
5. Walker ER, McGee RE, Druss BG. Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. JAMA Psychiatry 2015 Apr;72(4):334-341. [doi: 10.1001/jamapsychiatry.2014.2502] [Medline: 25671328]
6. Prince M, Patel V, Saxena S, et al. No health without mental health. Lancet 2007 Sep 8;370(9590):859-877. [doi: 10.1016/S0140-6736(07)61238-0] [Medline: 17804063]
7. Rentenversicherung in Zeitreihen [German pension insurance. Pension insurance in time series]. Deutsche Rentenversicherung. 2024. URL: https://www.deutsche-rentenversicherung.de/SharedDocs/Downloads/DE/Statistiken-und-Berichte/statistikpublikationen/rv_in_zeitreihen.html [accessed 2025-01-28]
8. Volkswirtschaftliche Kosten durch Arbeitsunfähigkeit 2023 [Federal institute for occupational safety and health. Economic costs of work disability 2023]. Bundesanstalt für Arbeitsschutz und Arbeitsmedizin. 2024. URL: https://www.baua.de/DE/Themen/Monitoring-Evaluation/Zahlen-Daten-Fakten/Kosten-der-Arbeitsunfaehigkeit [accessed 2025-01-28]
9. Health at a glance: Europe 2018. : OECD; 2018 URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2018/11/health-at-a-glance-europe-2018_g1g91fe4/health_glance_eur-2018-en.pdf [accessed 2025-01-28]
10. Wittchen HU, Jacobi F, Rehm J, et al. The size and burden of mental disorders and other disorders of the brain in Europe 2010. Eur Neuropsychopharmacol 2011 Sep;21(9):655-679. [doi: 10.1016/j.euroneuro.2011.07.018] [Medline: 21896369]
11. Wittchen HU, Jacobi F. Size and burden of mental disorders in Europe--a critical review and appraisal of 27 studies. Eur Neuropsychopharmacol 2005 Aug;15(4):357-376. [doi: 10.1016/j.euroneuro.2005.04.012] [Medline: 15961293]
12. World mental health report: transforming mental health for all. : World Health Organization; 2022 1st URL: https://iris.who.int/server/api/core/bitstreams/40e5a13a-fe50-4efa-b56d-6e8cf00d5bfa/content [accessed 2024-11-24]
13. Moitra M, Santomauro D, Collins PY, et al. The global gap in treatment coverage for major depressive disorder in 84 countries from 2000-2019: a systematic review and Bayesian meta-regression analysis. PLoS Med 2022 Feb;19(2):e1003901. [doi: 10.1371/journal.pmed.1003901] [Medline: 35167593]
14. Barbato A, Vallarino M, Rapisarda F, Lora A, Almeida J. EU compass for action on mental health and well-being—access to mental health care in Europe—a consensus paper. : European Union in the frame of the 3rd EU Health Programme (2014-2020); 2016 URL: https://health.ec.europa.eu/system/files/2016-12/ev_20161006_co04_en_0.pdf [accessed 2024-10-24]
15. Kruse J, Kampling H, Bouami SF, et al. Outpatient psychotherapy in Germany—an evaluation of the structural reform. Dtsch Arztebl Int 2024 May 17;121(10):315-322. [doi: 10.3238/arztebl.m2024.0039] [Medline: 38544323]
16. Tarricone R, Petracca F, Weller HM. Towards harmonizing assessment and reimbursement of digital medical devices in the EU through mutual learning. NPJ Digit Med 2024 Oct 1;7(1):268. [doi: 10.1038/s41746-024-01263-w] [Medline: 39354125]
17. Rodriguez-Villa E, Torous J. Regulating digital health technologies with transparency: the case for dynamic and multi-stakeholder evaluation. BMC Med 2019 Dec 3;17(1):226. [doi: 10.1186/s12916-019-1447-x] [Medline: 31801532]
18. Das Fast-Track-Verfahren für Digitale Gesundheitsanwendungen (DiGA) nach § 139e SGB V-Ein Leitfaden für Hersteller, Leistungserbringer und Anwender [The fast-track procedure for digital health applications (DHA) according to § 139e SGB

V. A Guide for Manufacturers, Healthcare Providers, and Users]. 2025 URL: https://www.bfarm.de/SharedDocs/Downloads/DE/Medizinprodukte/diga_leitfaden.html [accessed 2025-12-21]
19. DiGA-Verzeichnis [DHA directory]. BfArM. 2024. URL: https://diga.bfarm.de/de/verzeichnis [accessed 2024-10-09]
20. Dahlhausen F, Zinner M, Bieske L, Ehlers JP, Boehme P, Fehring L. Physicians' attitudes toward prescribable mHealth apps and implications for adoption in Germany: mixed methods study. JMIR Mhealth Uhealth 2021 Nov 23;9(11):e33012. [doi: 10.2196/33012] [Medline: 34817385]
21. Posselt J, Lander J, Dierks ML. Digitale Gesundheitsanwendungen in der hausärztlichen Versorgung: eine Diskussionsgrundlage zur Förderung informierter Nutzungsentscheidungen [Digital health applications in primary care: a basis for discussion to promote informed decisions on use]. Präv Gesundheitsf 2024 Nov;19(4):483-489. [doi: 10.1007/s11553-024-01126-y]
22. Gagnon MP, Desmartis M, Labrecque M, et al. Systematic review of factors influencing the adoption of information and communication technologies by healthcare professionals. J Med Syst 2012 Feb;36(1):241-277. [doi: 10.1007/s10916-010-9473-4] [Medline: 20703721]
23. DiGA-Report II 2024. Techniker Krankenkasse. 2024. URL: https://www.tk.de/resource/blob/2170850/e7eaa59ecbc0488b415409d5d3a354cf/tk-diga-report-2-2024-data.pdf [accessed 2024-10-11]
24. Bericht des GKV-Spitzenverbandes über die Inanspruchnahme und Entwicklung der Versorgung mit Digitalen Gesundheitsanwendungen (DiGA-Bericht)—Berichtszeitraum 2023 [Report of the GKV Spitzenverband on the Utilization and Development of Care with Digital Health Applications (DHA Report). Reporting Period 2023]. 2023 URL: https://www.gkv-spitzenverband.de/media/dokumente/krankenversicherung_1/telematik/digitales/2023_DiGA_Bericht_GKV-Spitzenverband.pdf [accessed 2024-11-24]
25. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
26. van Tilburg ML, Spin I, Pisters MF, et al. Barriers and facilitators to the implementation of digital health services for people with musculoskeletal conditions in the primary health care setting: systematic review. J Med Internet Res 2024 Aug 27;26:e49868. [doi: 10.2196/49868] [Medline: 39190440]
27. Kreuzenbeck CCJ, Schneider BS, Brenner SX, Koerber F. Rapid review on the incentives of digital health apps for physicians and psychotherapists: a German perspective. Digit HEALTH 2024;10:20552076241242781. [doi: 10.1177/20552076241242781] [Medline: 38698827]
28. Creswell JW, Plano Clark VL. Designing and Conducting Mixed Methods Research, 2nd edition: Sage; 2011.
29. Creswell JW. Qualitative Inquiry and Research Design: Choosing among Five Approaches, 3rd edition: Sage; 2013.
30. Bazeley P. Integrative analysis strategies for mixed data sources. American Behavioral Scientist 2012 Jun;56(6):814-828. [doi: 10.1177/0002764211426330]
31. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 7;4:50. [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
32. Field A. Discovering Statistics Using IBM SPSS Statistics, 5th edition: Sage; 2018.
33. Mergenthal K, Güthlin C, Klein AA. SaxoForN—Transregionales allgemeinmedizinisches Forschungspraxennetz Dresden und Frankfurt am Main: Konzept einer innovativen Forschungspraxeninfrastruktur [SaxoForN—transregional general practice research network Dresden and Frankfurt am Main: concept of an innovative research practice infrastructure]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2023;66(9):1042-1050. [doi: 10.1007/s00103-023-03722-3] [Medline: 37311815]
34. Gerber M, Schütze D, Dieckelmann M, Siebenhofer A, Engler J. Gesundheitsförderung in der Hausarztpraxis—eine qualitative Studie zur Rolle der Medizinischen Fachangestellten [Health promotion in general practice—a qualitative study on the role of medical assistants]. Z Allg Med 2023 Oct;99(6):291-295. [doi: 10.1007/s44266-023-00101-8]
35. Mergenthal K, Güthlin C, Beyer M, Gerlach F, Siebenhofer A. Wie bewerten und akzeptieren Patienten die Betreuung durch Medizinische Fachangestellte in der Hausarztpraxis? Ergebnisse einer Patienten-Befragung in der HzV in Baden-Württemberg [How do patients evaluate and accept care by medical assistants in general practice? Results of a patient survey in the hzv program in Baden-Württemberg]. Gesundheitswesen 2018 Dec;80(12):1077-1083. [doi: 10.1055/s-0042-110402]
36. Mayring P, Fenzl T. Qualitative inhaltsanalyse [Qualitative content analysis]. In: Baur N, Blasius J, editors. Handbuch Methoden Der Empirischen Sozialforschung [Handbook of Methods of Empirical Social Research]: Springer Fachmedien Wiesbaden; 2019:633-648. [doi: 10.1007/978-3-658-21308-4_42]
37. Cirkel L, Lechner F, Schlicker N, et al. Adoption and perception of prescribable digital health applications (DiGA) and the advancing digitalization among German internal medicine physicians: a cross-sectional survey study. BMC Health Serv Res 2024 Nov 6;24(1):1353. [doi: 10.1186/s12913-024-11807-1] [Medline: 39506735]
38. Byambasuren O, Beller E, Glasziou P. Current knowledge and adoption of mobile health apps among Australian general practitioners: survey study. JMIR Mhealth Uhealth 2019 Jun 3;7(6):e13199. [doi: 10.2196/13199] [Medline: 31199343]
39. Finley CR, Chan DS, Garrison S, et al. What are the most common conditions in primary care? Systematic review. Can Fam Physician 2018 Nov;64(11):832-840. [Medline: 30429181]

40. Jacob C, Sanchez-Vazquez A, Ivory C. Social, organizational, and technological factors impacting clinicians' adoption of mobile health tools: systematic literature review. JMIR Mhealth Uhealth 2020 Feb 20;8(2):e15935. [doi: 10.2196/15935] [Medline: 32130167]
41. O'Donnell A, Kaner E, Shaw C, Haighton C. Primary care physicians' attitudes to the adoption of electronic medical records: a systematic review and evidence synthesis using the clinical adoption framework. BMC Med Inform Decis Mak 2018 Nov 13;18(1):101. [doi: 10.1186/s12911-018-0703-x] [Medline: 30424758]
42. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146. [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
43. Wozney L, Newton AS, Gehring ND, et al. Implementation of eMental Health care: viewpoints from key informants from organizations and agencies with eHealth mandates. BMC Med Inform Decis Mak 2017 Jun 2;17(1):78. [doi: 10.1186/s12911-017-0474-9] [Medline: 28577543]
44. Schröder J, Berger T, Meyer B, et al. Attitudes towards internet interventions among psychotherapists and individuals with mild to moderate depression symptoms. Cogn Ther Res 2017 Oct;41(5):745-756. [doi: 10.1007/s10608-017-9850-0]
45. Lluch M. Healthcare professionals' organisational barriers to health information technologies-a literature review. Int J Med Inform 2011 Dec;80(12):849-862. [doi: 10.1016/j.ijmedinf.2011.09.005] [Medline: 22000677]
46. Wangler J, Jansky M. How can primary care benefit from digital health applications?—a quantitative, explorative survey on attitudes and experiences of general practitioners in Germany. BMC Digit Health 2024;2(1). [doi: 10.1186/s44247-024-00068-x]
47. Schwarz J, Chen K, Dashti H, et al. Piloting digital navigators to promote acceptance and engagement with digital mental health apps in German outpatient care: protocol for a multicenter, single-group, observational, mixed methods interventional study (DigiNavi). JMIR Res Protoc 2025 Sep 25;14:e67655. [doi: 10.2196/67655] [Medline: 40996088]
48. Wangler J, Jansky M. Welche Potenziale und Mehrwerte bieten DiGA für die hausärztliche Versorgung?—Ergebnisse einer Befragung von Hausärzt*innen in Deutschland [What potentials and added values do digital health applications (DiGA) offer for primary care?—Results of a survey of general practitioners in Germany]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2022 Dec;65(12):1334-1343. [doi: 10.1007/s00103-022-03608-w]
49. Scott A, Bai T, Zhang Y. Association between telehealth use and general practitioner characteristics during COVID-19: findings from a nationally representative survey of Australian doctors. BMJ Open 2021 Mar 24;11(3):e046857. [doi: 10.1136/bmjopen-2020-046857] [Medline: 33762248]
50. Holanda AA, do Carmo E Sá HL, Vieira A, Catrib AMF. Use and satisfaction with electronic health record by primary care physicians in a health district in Brazil. J Med Syst 2012 Oct;36(5):3141-3149. [doi: 10.1007/s10916-011-9801-3] [Medline: 22072279]
51. Spoont M, Greer N, Su J, Fitzgerald P, Rutks I, Wilt TJ. Rural vs. urban ambulatory health care: a systematic review. : Department of Veterans Affairs Health Services Research & Development Service; 2011 URL: https://www.ncbi.nlm.nih.gov/books/NBK56144/pdf/Bookshelf_NBK56144.pdf [accessed 2024-10-24]
52. Police R, Foster T, Wong K. Adoption and use of health information technology in physician practice organisations: systematic review. J Innov Health Inform ;18(4):245-258. [doi: 10.14236/jhi.v18i4.780]
53. Simon SR, Kaushal R, Cleary PD, et al. Correlates of electronic health record adoption in office practices: a statewide survey. J Am Med Inform Assoc 2007;14(1):110-117. [doi: 10.1197/jamia.M2187] [Medline: 17068351]
54. Pollack CE, Soulos PR, Herrin J, et al. The impact of social contagion on physician adoption of advanced imaging tests in breast cancer. J Natl Cancer Inst 2017 Aug 1;109(8):djw330. [doi: 10.1093/jnci/djw330] [Medline: 28376191]
55. Hilbert M. The end justifies the definition: the manifold outlooks on the digital divide and their practical usefulness for policy-making. Telecomm Policy 2011 Sep;35(8):715-736. [doi: 10.1016/j.telpol.2011.06.012]
56. Bol N, Helberger N, Weert JCM. Differences in mobile health app use: a source of new digital inequalities? The Information Society 2018 May 27;34(3):183-193. [doi: 10.1080/01972243.2018.1438550]
57. Equity within digital health technology within the WHO European region: a scoping review. : World Health Organization; 2022 URL: https://iris.who.int/bitstream/handle/10665/365326/WHO-EURO-2022-6810-46576-67595-eng.pdf?sequence=1 [accessed 2025-02-19]
58. Machleid F, Jansky B, Wild V, Wiegmann C, Kaminski J, Schreiter S. Mobile Gesundheitstechnologien für eine gerechte Versorgung bei psychischen Erkrankungen [Mobile health technologies for equitable care in mental disorders]. Nervenheilkunde 2024 Nov;43(12):688-700. [doi: 10.1055/a-2415-8433]

Abbreviations

AI: artificial intelligence
BfArM: German Federal Institute for Drugs and Medical Devices (German: Bundesinstitut für Arzneimittel und Medizinprodukte)
DHA: digital health application
DHA-MD: digital health application for mental disorder

GP: general practitioner
MA: medical assistant
mHealth: mobile health

Edited by J Torous; submitted 06.Jun.2025; peer-reviewed by J Zielasek, J Schwarz; revised version received 28.Oct.2025; accepted 07.Nov.2025; published 06.Jan.2026.

Please cite as:
Scheibe S, Salm S, Mergenthal K, Engesser D, Stalujanis E, Singer S, Kemmerer P, Dotzauer L, Voigt K
General Practitioners' Perspectives on Digital Health Applications for Mental Disorders and Their Prescribing Behavior: Mixed Methods Study
JMIR Ment Health 2026;13:e78659
URL: https://mental.jmir.org/2026/1/e78659
doi:10.2196/78659

Blended Therapy From the Perspective of Mental Health Professionals in Routine Mental Health Care: Mixed Methods Analysis of Cross-Sectional Survey Data

Annalena Maria Kneubühler, MSc; Elianne von Känel, MSc; Kristina Grgic, MSc; Ena Munkovic, MSc; Thomas Berger, PhD; Laura Luisa Bielinski, PhD

Department of Clinical Psychology and Psychotherapy, Institute of Psychology, University of Bern, Fabrikstrasse 8, Bern, Switzerland

Corresponding Author:
Laura Luisa Bielinski, PhD
Department of Clinical Psychology and Psychotherapy, Institute of Psychology, University of Bern, Fabrikstrasse 8, Bern, Switzerland

Abstract

Background: Digital interventions play an innovative role in the treatment of mental health disorders, offering evidence-based solutions across a wide range of conditions. Blended therapy (BT), which integrates digitally delivered interventions with face-to-face therapy, has shown promise. However, challenges such as low uptake hinder widespread implementation. Mental health professionals are key stakeholders for the adoption of BT in routine care settings.

Objective: This study explores mental health professionals' perspectives on BT, specifically assessing their perceived knowledge of, acceptance of, usage of, and perceptions of different BT types. Additionally, it examines mental health professionals' perceived advantages and disadvantages of BT, challenges associated with implementation, and wishes toward the future application of BT.

Methods: A survey study was conducted among 203 mental health professionals (152 psychological psychotherapists and 51 psychiatrists, including also individuals in training) in Switzerland. The data were analyzed using both quantitative methods and qualitative content analysis.

Results: Participants reported limited knowledge of BT (mean 2.71, SD 1.32), attitudes toward BT were somewhat positive (mean 5.25, SD 1.34), and acceptance was moderate (mean 3.64, SD 1.20). Among various digitally delivered interventions, teletherapy (video) was most frequently integrated with face-to-face treatment and considered more suitable for BT than chat, email, or new technologies. More than 75% (n=152) of the respondents deemed BT appropriate for the treatment of affective (mood) disorders (F30-F39) and for the treatment of neurotic, stress-related, and somatoform disorders (F40-F48; ICD-10). The qualitative analyses of open-ended questions highlighted key advantages of BT as perceived by mental health professionals. These include increased treatment flexibility, the ability to outsource therapy components, and enhanced treatment efficiency. However, disadvantages such as increased effort and potential disruptions to the therapeutic relationship were also noted. Participants identified barriers to BT implementation, including financing and data security concerns. To facilitate BT adoption, respondents emphasized the desire for better cost coverage, easy access to digitally delivered interventions, and seamless integration of digital tools into face-to-face therapy.

Conclusions: The findings indicate that mental health professionals report limited knowledge of BT and consider it more suitable for certain disorders than others. Moreover, from their perspective, while BT offers advantages, it also presents disadvantages. Addressing mental health professional knowledge gaps, alongside resolving perceived implementation barriers, may be key to the successful future implementation of BT in routine mental health settings.

(JMIR Ment Health 2026;13:e78079) doi:10.2196/78079

KEYWORDS

blended therapy; digital interventions; implementation; routine care; therapist attitudes

Introduction

Digital Interventions to Treat Mental Health Disorders

The evidence base supporting the efficacy of digital psychological interventions to treat mental health conditions and problems is extensive and continues to grow [1-5]. These interventions encompass a broad spectrum, ranging from fully self-guided programs [4], designed to, for example, provide support for individuals who might otherwise lack access to traditional therapy, to more integrated approaches that combine digital elements with face-to-face therapy [6,7].

Blended Therapy

Blended therapy (BT) in the mental health context refers to the combination of a digitally delivered intervention and face-to-face therapy [8,9]. Digital interventions and face-to-face therapy can be combined in different ways, ranging from digitally delivered interventions provided prior to or as aftercare to face-to-face therapy to interventions interwoven during a course of face-to-face therapy. The first systematic review on BT [9] describes the potential of this type of treatment regarding both study dropout and time savings in therapy. A more recent systematic review and meta-analysis [2] describes the feasibility of BT and reports on BT effects. BT interventions were more effective or noninferior to treatment as usual (defined as pharmacological or psychological intervention and standard medical care), with a moderate-to-large effect size in the treatment of depression (Cohen d=–1.1, 95% CI –0.6 to –1.6; P<.001). For anxiety outcomes, the meta-analysis reported a small, nonsignificant effect size (Cohen d=–0.1, 95% CI –0.3 to 0.05; P=.17). The findings also highlight higher effect sizes for blended interventions with supplementary design, fewer (≤6) face-to-face sessions, and a lower ratio (≤50%) of face-to-face versus digital sessions [2].

BT in Routine Mental Health Care Settings

Various studies highlight the successful integration of digital interventions with face-to-face therapy in routine mental health care settings. Reported benefits include enhanced efficacy and effectiveness [6,10]. Another study reported no significant difference in symptom change over time between the blended and control group [11]. Moreover, research also underscores challenges and limitations associated with BT. For instance, a recent large-scale study conducted in routine care settings in Germany by Schaeuffele et al [12] identified issues such as adherence as hindering factors for implementation.

The Perception of Health Care Providers

Mental health care providers play a pivotal role in the successful implementation of BT. Their attitudes can influence practical application [13,14]. While lagging implementation of digitally delivered interventions appears to be a recurring trend across multiple European countries [14,15], the COVID-19 pandemic has led to greater uptake and acceptance [16,17]. BT has been perceived more favorably than stand-alone digitally delivered interventions by clinicians [13,18,19]. However, reservations toward BT among mental health professionals have also been reported. For example, concerns regarding the therapeutic alliance, patient engagement, data security, the therapeutic process, and work-life balance [20,21] may impact providers' willingness to adopt BT.

Aims

Although the evidence base on BT is growing, several research gaps remain. Most existing studies have focused on feasibility and clinical outcomes, while less is known about how BT is perceived and implemented in routine mental health care settings. Detailed insights into health professionals' perceived knowledge, attitudes, and acceptance of BT in Switzerland are limited, and both qualitative and quantitative analyses are required to adequately examine these specific topics. This study reports on a mixed methods analysis using data from a survey completed by mental health professionals and mental health professionals in training in Switzerland. Specifically, the study explores the following research questions: (1) What is the current level of perceived knowledge, attitude toward, and acceptance of BT among psychological psychotherapists and psychiatrists (including those in training)? (2) How is BT currently used by participants? (3) How do mental health professionals perceive the suitability of different digitally delivered interventions for BT purposes, and which types of BT are they willing to use in the future? (4) What are the perceived advantages and disadvantages of BT, what challenges are there regarding implementation and what are mental health professionals' wishes for the future regarding BT?

Methods

Study Design

This study examined BT from the perspective of psychological psychotherapists and psychiatrists (also those in training) in Switzerland, using a cross-sectional, open online-survey approach. Participants filled out the survey between October 2023 and February 2024.

Ethical Considerations

The study received approval from the Ethics Commission of the Faculty of Human Sciences, University of Bern (ID: 2023-09-04). Participants received no incentive or compensation for participation. All participants provided informed consent to participate. The survey was conducted with no collection of direct identifiers such as names, contact information, IP addresses, or geographic location. The survey included limited demographic variables (eg, gender and job category) for analytical purposes. Any potentially identifying information contained in free-text responses was removed or generalized prior to analysis.

Measures

A total of 23 survey questions from a comprehensive survey on the topic of BT were used to answer the research questions presented in this study. The full survey translated from German to English can be found in Multimedia Appendix 1 along with the instructions participants received. Survey questions were built on previous literature [8,14,18-20,22,23]. The survey was provided through Qualtrics [24] and was tested prior to dissemination with several test-runs by the authors of this study. Users' IP addresses were not recorded. The survey was available in German and French for participants. Each survey page included a back button.

To answer research question 1, we assessed mental health professionals' perceived knowledge of, general attitude toward, and acceptance of BT. Acceptance of BT was operationalized following Braun et al [22] using 3 specific items: "I could imagine including BT into my work"; "I intend to try out BT in my work within the next year"; "How high is your intention to use BT in your work ever?". The first 2 questions were assessed on a 5-point Likert scale ranging from 1 (totally disagree) to 5 (totally agree). The third item was rated on a 0-to-100 scale and converted into a 5-point Likert scale to measure the strength of intention. A mean value was calculated from all 3 items to quantify the acceptance of BT. Based on prior research [22], the mean acceptance score was categorized as low (1 - 2.34), moderate (2.35 - 3.67), or high (3.68 - 5). To answer research question 2, we assessed both past use of BT and current use of the different digital intervention modalities for BT (eg, teletherapy [video], chat, email, self-management, new technologies). To answer research question 3, the perceived suitability of different digitally delivered interventions (teletherapy [video], chat, email, self-management interventions, and new technologies) for BT was assessed. The suitability of BT for different *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) [25] disorders was also assessed. Moreover, the future willingness to use digital interventions in relation to various points of treatment and in different settings (outpatient, day clinic, inpatient, acute inpatient) was assessed. To answer research question 4, participant answers to 4 open-ended questions were examined. Detailed item wording and the precise response scales for all items used to answer the research questions are reported in Multimedia Appendix 2.

Statistical Analyses

All quantitative analyses were conducted using SPSS (version 29; IBM Corp). Descriptive statistics (means, SDs, frequencies, and percentages) were used to address the primary research questions. Inferential statistics were applied to explore patterns and group differences. Repeated-measures ANOVAs were conducted where appropriate, with Greenhouse-Geisser corrections applied when the assumptions of sphericity were violated. For participants with missing values, listwise deletion was applied. Effect sizes ηp^2 were reported to aid interpretation for ANOVAs. Pairwise comparisons were Bonferroni corrected and Cohen *dz* was reported as effect size. For dichotomous outcomes, Cochran *Q* and follow-up McNemar tests with Bonferroni corrections were used, and Cohen *g* was reported as effect size for the pairwise comparisons. For group comparisons between professional groups and between those in training versus not in training, independent sample *t* tests were conducted. All significance tests were 2-sided with a significance level of α=.05.

The perceived advantages and disadvantages of BT, as well as implementation challenges and future wishes, were analyzed using an inductive content analysis approach as outlined by Mayring [26]. This approach is well suited to qualitative analyses that stay close to the semantic content of responses and allow for integration of qualitative and quantitative elements, such as reporting category frequencies. Separate inductive analyses were conducted for each area (advantages, disadvantages, challenges, and future wishes). Following Mayring's [26] category formation steps, KG coded all responses, assigning codes to the raw material. Multiple codes could be assigned per survey item response, but the same code could not be assigned twice. In the next step, categories and subcategories were discussed collaboratively with LLB, and the category system was refined in an iterative process of repeated reviewing of the material and adjusting of categories. Finally, KG coded the entire material set with the final categories and subcategories that were formed. All analyses were conducted using Microsoft Excel (version 2016). Anchor examples for the categories were taken verbatim from participant answers.

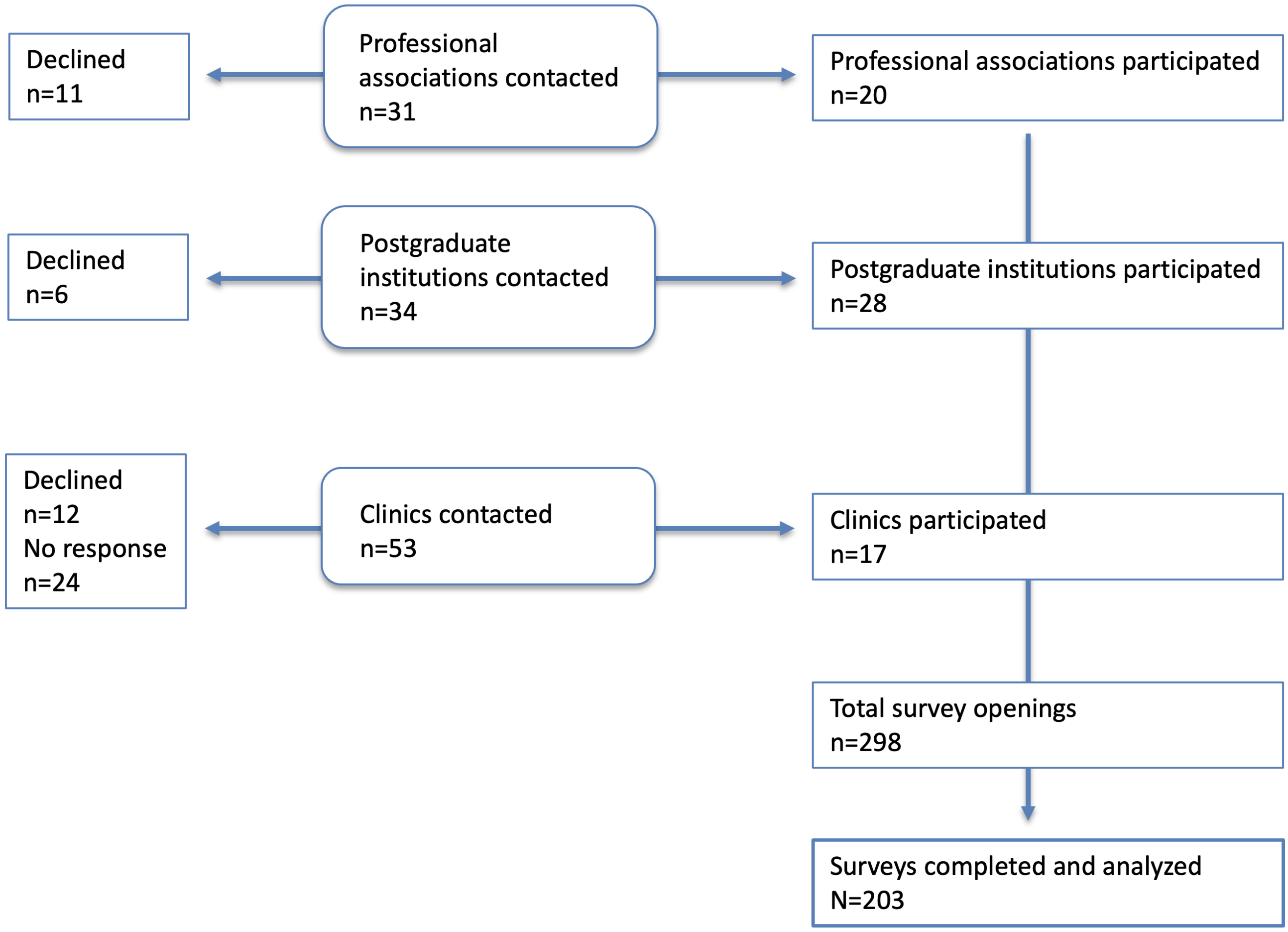
Due to variation in response rates across survey items, sample sizes are reported throughout the study. Detailed information on item-level missingness is provided in Multimedia Appendix 3. No weighting of items or propensity scores was used to adjust for the nonrepresentative sample.

Results

Recruitment

To recruit participants, professional associations, psychotherapy training institutes, and psychiatric clinics across Switzerland were contacted and invited to disseminate the study link to their members or personnel via internal communication channels. Up to 3 reminder emails were sent to each organization. The contacted clinics were identified from a public registry provided by the Schweizerisches Institut für ärztliche Weiter- und Fortbildung [27]. Overall, a broad range of professional and institutional stakeholders were approached, of whom a subset actively declined participation due to staff shortages, an overload of inquiries, or other individual reasons. A detailed overview of the recruitment process, including the number of institutions contacted and participating, is presented in Figure 1. The survey was opened 298 times, and the 203 responses that reached the end of the survey were included in the analysis.

Figure 1. Flow chart depicting recruitment pathways for the survey.



Sample

An overview of sample characteristics provided for the 203 survey completions (visited each survey page until the end) is presented in Table 1.

Table . Sample characteristics.

Sample characteristic	Values
Gender, n (%)	
Man	61 (30.0)
Woman	141 (69.5)
Nonbinary (diverse)	1 (0.5)
Age (y), mean (SD; Min[a], Max[b])	45.9 (14.1; 24, 79)
Professional group, n (%)	
In training to become a federally recognized psychotherapist	61 (30.0)
Federally recognized psychotherapist	91 (44.8)
Specialist in psychiatry and psychotherapy	41 (20.2)
In training to become a specialist in psychiatry and psychotherapy	7 (3.4)
Specialist in child and adolescent psychiatry and psychotherapy	3 (1.5)
Years of training, n (%)[c]	
First year	17 (25.0)
Second year	19 (27.9)
Third year	11 (16.2)
Fourth year	12 (17.6)
Fifth year	6 (8.8)
Sixth year	1 (1.5)
>6	2 (2.9)
Work experience in psychotherapeutic practice (y), n (%)	
None	1 (0.5)
<1	12 (5.9)
1 - 5	62 (30.5)
6 - 10	22 (10.8)
11 - 15	32 (15.8)
>15	74 (36.5)
Therapeutic orientation, n (%)[d]	
Cognitive-behavioral therapy (cognitive or cognitive-behavioral approach)	112 (55.2)
Depth-psychological or psychodynamic	37 (18.2)
Psychoanalytic	32 (15.8)
Systemic	63 (31.0)
Humanistic	48 (23.6)
Other	44 (21.7)
Current work setting, n (%)	
Outpatient	144 (70.9)
Partial inpatient or day clinic	3 (1.5)
Inpatient	22 (10.8)
Mixed (outpatient and inpatient)	19 (9.4)
Mixed (outpatient and partial inpatient)	6 (3.0)
Mixed (partial inpatient and inpatient)	6 (3.0)
Currently not employed	3 (1.5)

[a]Min: minimum.

[b]Max: maximum.

[c]This applies to the subgroups in training to become a federally recognized psychotherapist and in training to become a specialist for psychiatry and psychotherapy.

[d]Multiple responses were possible.

Perceived Knowledge of, Attitude Toward, and Acceptance of BT

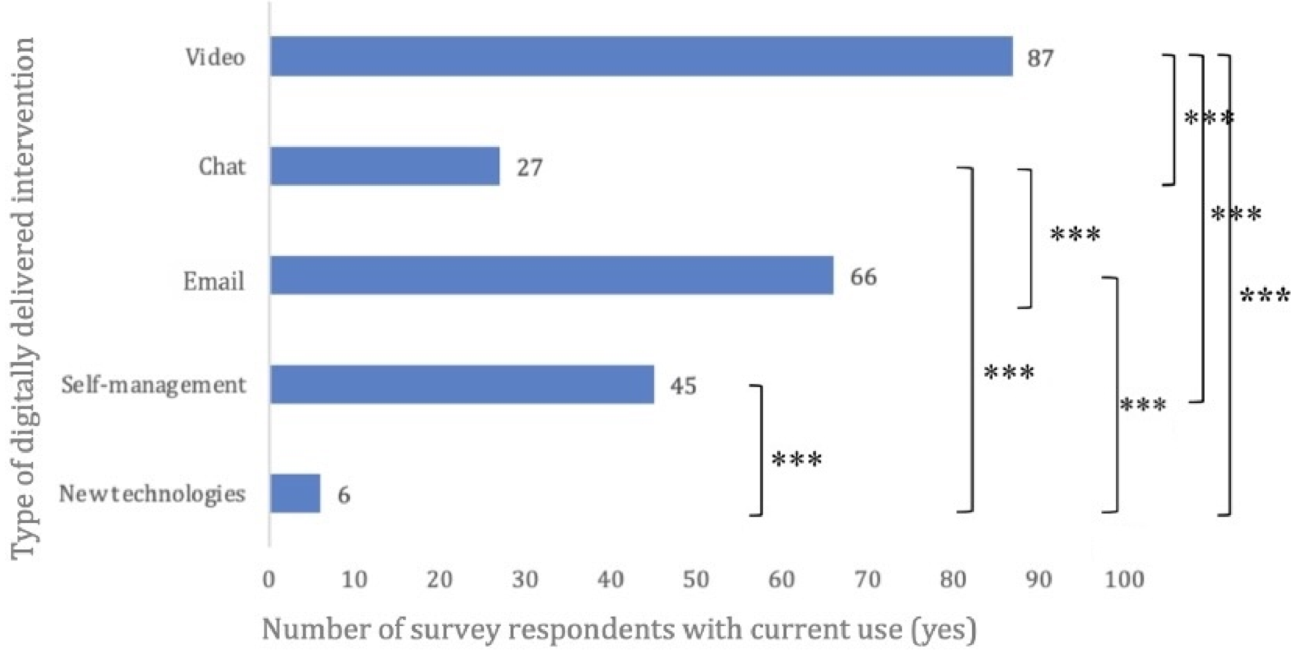
The overall sample reported a mean (SD) of 2.71 (1.32) for perceived knowledge, corresponding to a value of 3 ("a little"). A total of 44 (21.7%) participants reported having no knowledge of BT, and only 4 (2.0%) participants reported having a great deal of knowledge of BT. See Table S1 in Multimedia Appendix 4 for full descriptive data. Regarding attitude toward BT, the overall sample reported a mean (SD) of 5.25 (1.34), corresponding to a value of 5 ("somewhat positive"). For BT acceptance, the mean (SD) was 3.64 (1.20), corresponding to moderate acceptance [22]. Analyses of differences in knowledge of BT, attitude toward BT, and acceptance between professional groups and between those in training versus those not in training are provided in Tables S2-S4 in Multimedia Appendix 4.

Use of BT

Of the total sample, 125 (61.6%) participants reported having used some form of BT in the past. The mean score for current use across the sample was mean (SD) of 2.14 (1.22), corresponding to a scale value of 2 ("rarely"). Figure 2 shows the number of participants who answered "yes" to currently using different types of digitally delivered interventions as part of therapy. A Cochran *Q* test indicated significant differences in current use across intervention types (Q_4 =136.58; N=203; *P*<.001). Pairwise McNemar tests were conducted to further examine these differences. To control for type I error inflation due to multiple comparisons (*k*=10), a Bonferroni correction was applied. After Bonferroni correction, all differences between digitally delivered intervention formats remained significant except for chat versus self-management (*P*=.21), self-management versus email (*P*=.12), and email versus video (*P*=.09). See Table S5 in Multimedia Appendix 4 for a full overview of the pairwise comparisons.

Comparisons between professional groups and those in training versus not in training regarding current use of BT are also shown in Tables S6 and 7a and 7b in Multimedia Appendix 4.

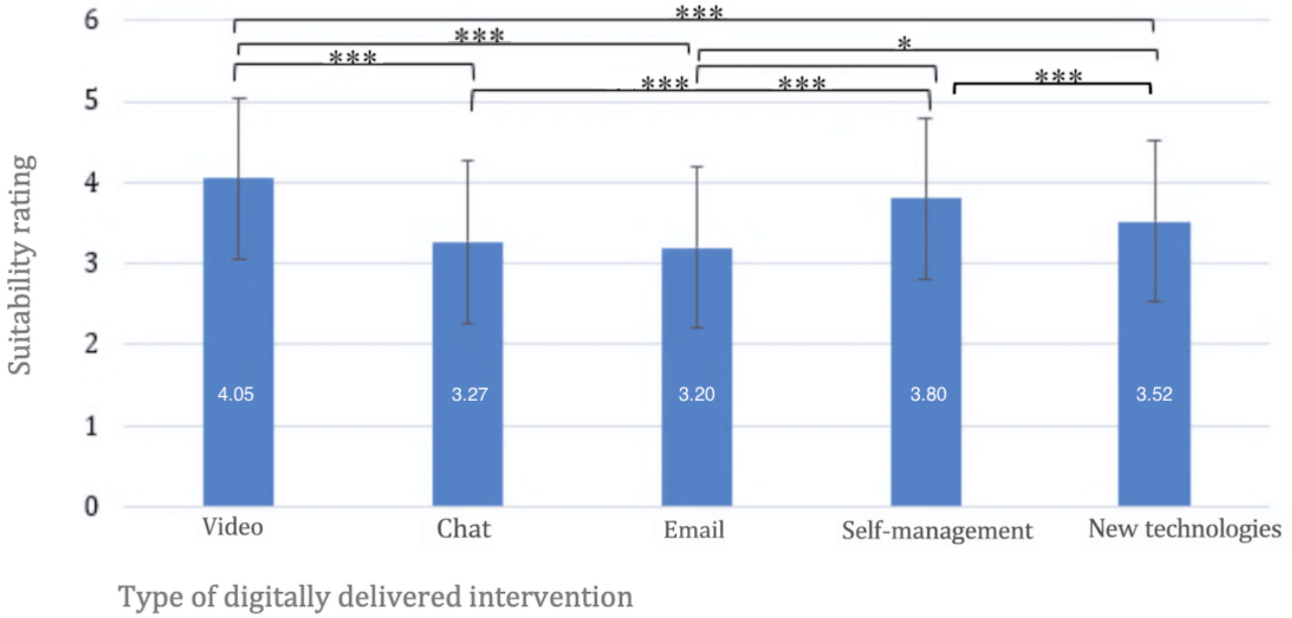
Figure 2. Current use of different digitally delivered interventions in combination with face-to-face therapy. N=203. ****P*<.001. The number of participants who answered *yes* to current use of different digitally delivered interventions in combination with face-to-face therapy by intervention type is displayed. Multiple digitally delivered intervention uses were possible per participant.



Perceived Suitability of Different Digitally Delivered Interventions for BT

Participants rated the suitability of different digitally delivered intervention types for BT. A repeated-measures ANOVA showed significant differences in suitability ratings between the intervention types ($F_{3.48,\ 702.27}$=30.57; *P*<.001; ηp^2=0.13). Video conferencing was rated as significantly more suitable than interventions via chat (mean difference [MD]=0.78; *P*<.001; *dz*=0.57), email (MD=0.86; *P*<.001; *dz*=0.60), and new technologies (MD=0.54; *P*<.001; *dz*=0.43). Chat was rated significantly less suitable than self-management interventions (MD=−0.53; *P*<.001; *dz*=0.39). Email interventions were rated significantly less suitable than self-management interventions (MD=−0.60; *P*<.001; *dz*=0.47) and new technologies (MD=−0.32; *P*=.01; *dz*=0.23); see also Figure 3. Table S8 in Multimedia Appendix 4 shows an overview for full descriptive data.

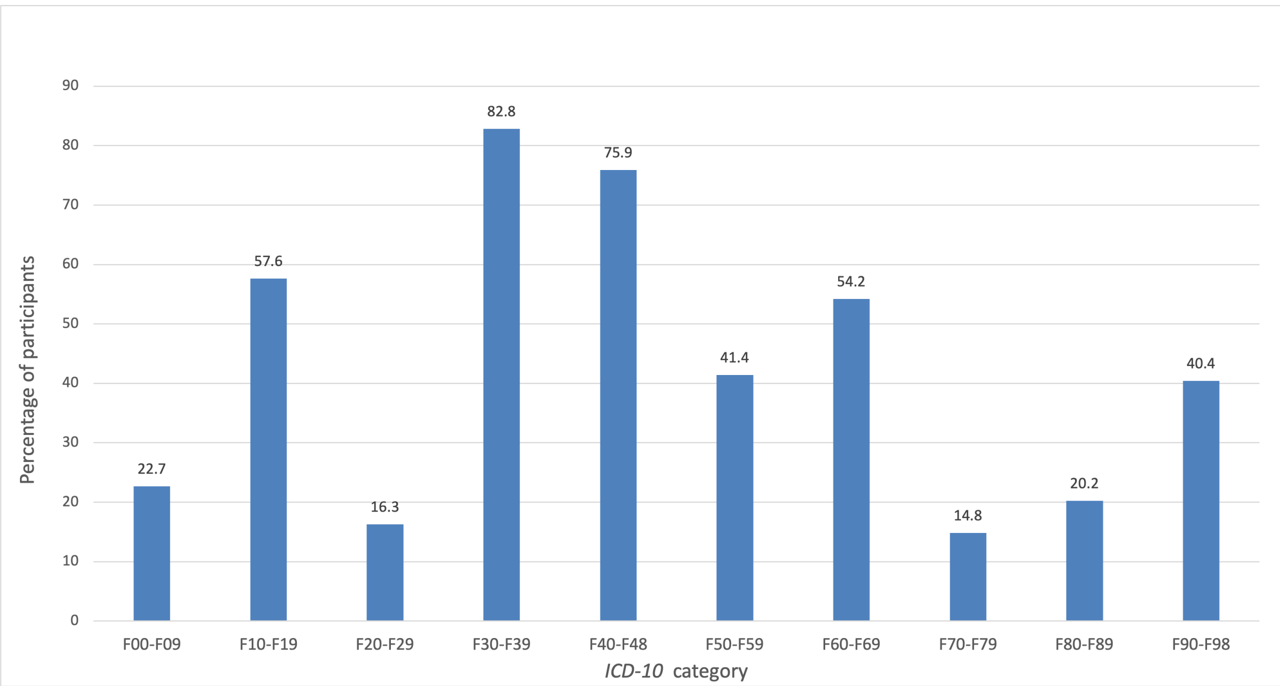
Figure 3. Suitability ratings of different digitally delivered interventions for blended therapy (BT). N=203. *P<.05; **P<.01; ***P<.001. Means are displayed in white. Error bars represent ±1 SD. Axis extends beyond the maximum response option to display full error bars; no respondent values exceeded the upper scale limit (5).



Suitability According to *ICD-10* Mental Health Disorder Categories

As shown in Figure 4, over 75% (n=152) of the participants considered BT suitable for treating Mood disorders (F30-F39) and Neurotic, stress-related, and somatoform disorders (F40-F48). In contrast, very few participants endorsed BT as suitable for Schizophrenia and delusional disorders (F20-F29) or Intellectual disabilities (F70-F79). The Cochran Q test indicated significant differences in suitability across disorder categories ($Q(9)=558.55$; $P<.001$). Pairwise McNemar tests with Bonferroni correction further explored these differences (see Table S9 in Multimedia Appendix 4).

Figure 4. Descriptive data on participants who rated blended therapy (BT) as suitable for different mental health disorders. N=203. Percentage of participants who said BT was suitable is displayed for each disorder group. F00-F09=Organic, including symptomatic, mental disorders; F10-F19=Mental and behavioral disorders due to psychoactive substance use; F20-F29=Schizophrenia, schizotypal, and delusional disorders; F30-F39=Mood [affective] disorders; F40-F48=Neurotic, stress-related, and somatoform disorders; F50-F59=Behavioral syndromes associated with physiological disturbances and physical factors; F60-F69=Disorders of adult personality and behavior; F70-F79=Intellectual disabilities; F80-F89=Disorders of psychological development; F90-F98=Behavioral and emotional disorders with onset usually occurring in childhood and adolescence. *ICD-10*: *International Statistical Classification of Diseases, Tenth Revision*.

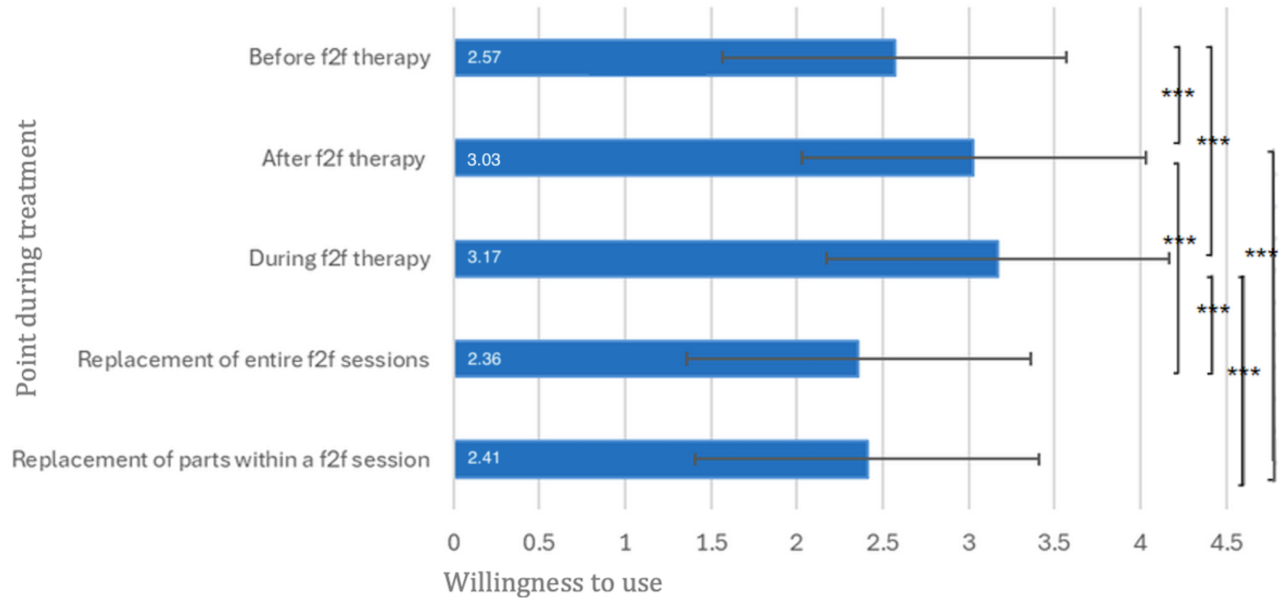


Willingness to Use at Different Points During Treatment

The future willingness to use digitally delivered interventions in relation to various points of treatment is illustrated in Figure 5. The mean (SD) values were 2.57 (1.06) for use before psychotherapy, 3.03 (0.89) for after psychotherapy, and 3.17 (0.84) for during psychotherapy. These values each correspond to scale point 3 (“rather yes”). For use as a substitute for individual sessions, the mean (SD) was 2.36 (1.12), and for use as a substitute for individual parts of a session, the mean (SD) was 2.41 (1.03), which in both cases corresponds to scale point 2 (“rather no”).

A repeated-measures ANOVA indicated significant differences between the application points ($F_{3.11,\ 591.20}$=41.34; P<.001; ηp^2=0.18). Willingness to use digitally delivered interventions was significantly lower before psychotherapy than after (MD=−0.46; P<.001; dz=0.51) and during psychotherapy (MD=−0.60; P<.001; dz=0.54). Willingness to use digitally delivered interventions after psychotherapy was significantly higher than for the replacement of individual sessions (MD=0.67; P<.001; dz=0.51) and for the replacement of individual parts of sessions (MD=0.62; P<.001; dz=0.63). Similarly, willingness to use digitally delivered interventions during psychotherapy was significantly higher than for the replacement of individual sessions (MD=0.81; P<.001; dz=0.68) and for the replacement of individual parts of sessions (MD=0.76; P<.001; dz=0.75). No other pairwise differences were statistically significant (see Figure 5).

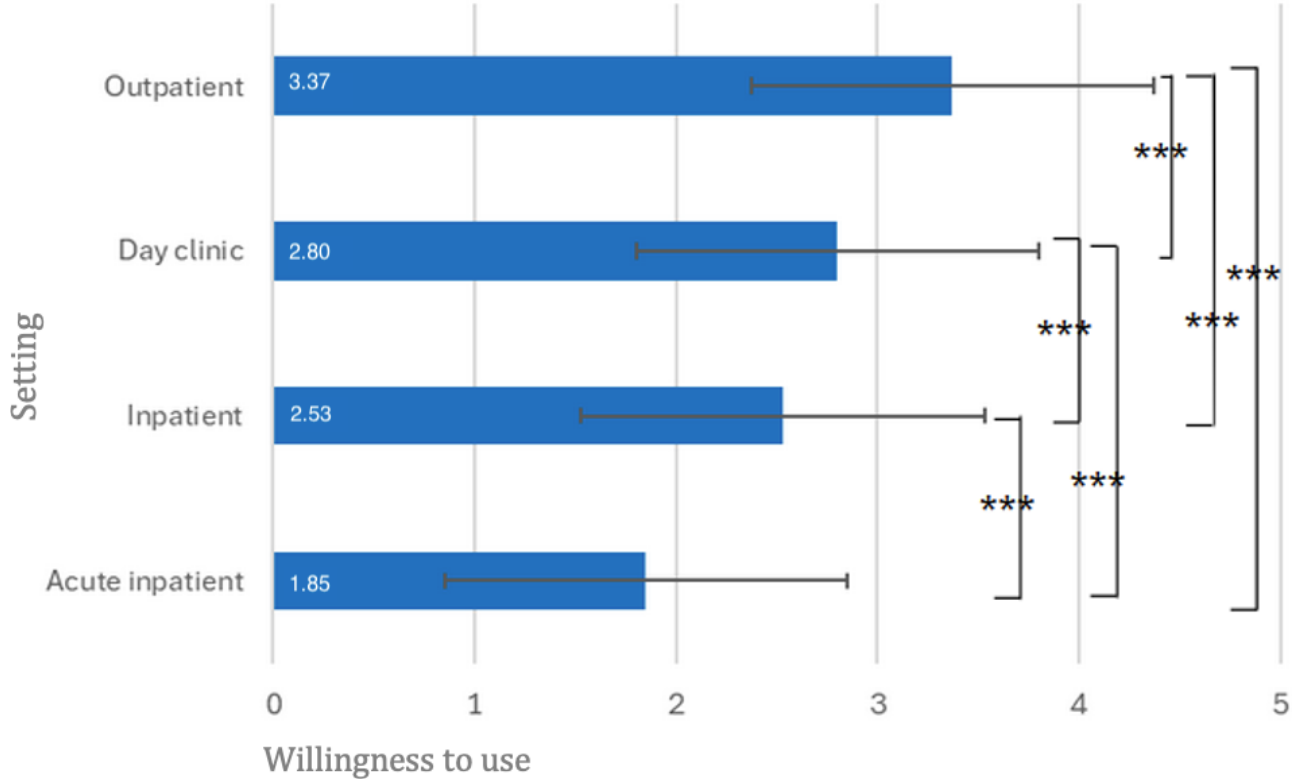
Figure 5. Future willingness to use digitally delivered interventions at different points during treatment. N=191. ***P<.001. Definitely no=1, Rather no=2, Rather yes=3, and Definitely yes=4. Means are displayed in white. Error bars represent ±1 SD. Axis extends beyond the maximum response option to display full error bars; no values exceeded the upper scale limit (4). f2f: face to face.



Willingness to Use Across Different Settings

The future willingness to use digitally delivered interventions varied across treatment settings (Figure 6). The mean (SD) ratings were 3.37 (0.72) for outpatient settings, 2.53 (0.88) for inpatient settings, 2.80 (0.89) for day clinic settings, and 1.85 (0.82) for acute inpatient settings. A repeated-measures ANOVA showed significant differences in willingness to use between the settings ($F_{2.75,\ 471.00}$=185.79; P<.001; ηp^2=0.52). Willingness to use was significantly lower in the acute inpatient setting than in the inpatient (MD=−0.68; P<.001; dz=0.84), day clinic (MD=−0.94; P<.001; dz=1.06), and outpatient (MD=−1.52; P<.001; dz=1.53) settings. It was also significantly lower for inpatient than day clinic (MD=−0.26; P<.001; dz=0.36) and outpatient (MD=−0.84; P<.001; dz=0.95) settings and significantly lower for day clinic than outpatient settings (MD=−0.58; P<.001; dz=0.71; see Figure 6).

Figure 6. Future willingness to use blended therapy (BT) in different settings. N=172. ***P<.001. Definitely no=1, Rather no=2, Rather yes=3, and Definitely yes=4. Means are displayed in white. Error bars represent ±1 SD. Axis extends beyond the maximum response option to display full error bars; no values exceeded the upper scale limit



Advantages and Disadvantages of BT

A total of 148 participants reported on the advantages of BT (Table 2). A total of 233 codes were coded. These were grouped into 4 main categories and 14 subcategories. At least 141 participants reported on the disadvantages of BT (Table 3). A total of 215 codes were generated, which were grouped into 5 main categories and 23 subcategories.

Table . Advantages of blended therapy (BT).

Category	Subcategory	n[a]	Anchor example
Therapy-related factors	Flexibility	79	Flexibly being able to cater to individuality of patients with therapy (Survey 5)
	Outsourcing therapy elements	43	Delegation of in-depth psychoeducation (Survey 101)
	Efficiency	35	More effective and thus shorter therapy duration (Survey 26)
	For different points in treatment	10	Bridge wait times (Survey 141)
	Problem-specific or transdiagnostic benefits	8	Autism/Asperger (Survey 37)
Relationship factors	Strengthening the therapeutic relationship	13	Strengthening of commitment and compliance in therapy (Survey 129)
Patient factors	Increase in self-efficacy	12	Increasing self-efficacy of patients (Survey 29)
	Lowering barriers	6	Reducing barriers to do with fear (Survey 17)
	Access to topics	5	Access to topics that patients can't talk about in face-to-face sessions (Survey 81)
	Increase in therapy motivation	4	Change in therapy motivation is measurable (Survey 129)
	Attractive option for therapy	2	Making therapy more attractive (Survey 37)
Mental health professional factors	Capacity	9	More clients (Survey 66)
	Location-independent work	4	Work in home office (Survey 26)
	Relief	3	Relief (Survey 97)

[a]n: number of participants with code.

Table . Disadvantages of blended therapy (BT).

Category	Subcategory	n[a]	Anchor example
Technological, organizational, and legal aspects	Technical prerequisites	15	Technology needs to work (Survey 170)
	Billing	11	Unclear billing options (Survey 31)
	Costs	10	Investment in further training and procurement (Survey 82)
	Data security	12	Need to engage with data security topic (Survey 6)
Practical implementation of interventions	Effort	24	Additional effort to gather information and review suitable options (Survey 40)
	Limitation of holistic treatment	15	The therapeutic vessel may become watered down (Survey 139)
	Not suitable for some interventions	9	Behavioral observation becomes more difficult (Survey 45)
	Progress monitoring	5	Less control over the development of the patients' condition (Survey 137)
	Competition among offers	1	Competition between individual offerings (Survey 33)
Interpersonal interaction	Therapeutic relationship	21	The relationship is interrupted (Survey 37)
	Reduced contact	11	Reduction of human contact (Survey 65)
	Nonverbal communication	3	Lack of nonverbal communication (Survey 132)
Patient-related challenges	Indication	24	My patients with psychosis will often not use it (Survey 145)
	Avoidance behavior	13	Enables avoidance of interactions with others (Survey 37)
	Lack of motivation	4	Patients' motivation is rather unclear and uncertain (Survey 61)
	Overwhelm	3	Overwhelm of the patient (Survey 65)
	Loss of autonomy	2	Misunderstandings in autonomous work (Survey 187)
	Media consumption	1	Dependency on tools (Survey 65)
Personal and professional challenges	Lack of knowledge	10	The topic is unclear to me: too little experience (Survey 106)
	Accessibility	8	Having to organize when one is not reachable, etc (Survey 96)
	Cognitive strain	4	Increased tiredness for therapist when contact is not face-to-face (Survey 96)
	Professional field	4	Loss of individuality and spontaneity in individual cases, emergence of boredom even for the therapist (Survey 173)

[a]n: number of participants with code.

Challenges for Implementation and Wishes for the Future

A total of 129 individuals were included in the qualitative analysis of open-ended responses to the question on challenges regarding BT implementation. A total of 7 main categories and 15 subcategories were identified. At the subcategory level, a total of 206 codes were assigned, as shown in Table 4. Additionally, 108 individuals were included in the analysis of the open-ended question on wishes for the future regarding BT. A total of 151 codes were generated and grouped into 15 categories, which are presented in Table 5.

Table . Challenges regarding implementation.

Category	Subcategory	n[a]	Anchor example
Technical challenges	Data security	25	Violating patient and data privacy (Survey 2)
	Usability	17	Difficulties in usage (Survey 8)
	Software and hardware	12	Development of good software (Survey 26)
Costs and financing	Direct costs	34	Financing (Survey 138)
	Indirect costs	20	Time investment (Survey 48)
Therapeutic relationship and quality of therapy	Relationship	11	Difficulty relationship building (Survey 37)
	Quality of therapy	9	Tendency toward superficiality (Survey 19)
Adaptability and flexibility	Choice of digitally delivered interventions	15	How do I know for example, if an app is good? (Survey 173)
	Individualization	8	All of therapy needs to be adaptable to the patient (Survey 9)
Motivation and acceptance	Patients	14	Skepticism for example amongst older patients (Survey 22)
	Mental health professionals	11	Acceptance amongst mental health professionals (Survey 183)
Training and knowledge	Training	8	Further training is necessary (Survey 32)
	Knowledge and familiarity	7	Too little knowledge about digitally delivered interventions (Survey 178)
Indication and suitability	Contraindication	10	Not during crises (Survey 176)
	Judging risk	5	Risk of missing signs of suicidality (Survey 5)

[a]n: number of participants with code.

Table . Wishes for the future regarding blended therapy (BT).

Category	n[a]	Anchor example
Costs being covered	16	Costs covered by insurance (Survey 100)
Easy access	16	More easily accessible and nationally available offers (Survey 62)
Easy integration into therapy	16	BT as a self-evident part of the psychotherapeutic treatment (Survey 7)
Use as an add-on	13	Only as a supplement to face-to-face therapy (Survey 75)
Knowledge provision	12	More education and knowledge about it (Survey 130)
Individual tailoring	12	Good options that can be adapted by both mental health professional and patient (Survey 35)
Specialized programs	10	Diagnosis-specific implementation (Survey 37)
Software development	9	Good programs. I have tried Velibra, which I find very good (Survey 31)
Flexibility in use	9	Therapeutic freedom (Survey 42)
Studies on efficacy	9	Studies on effectiveness of digitally delivered interventions (Survey 182)
Training	8	Practice-based training (Survey 8)
Secure use	5	Moderately and with mindfulness towards the protection of personality (Survey 19)
Increased acceptance	4	More willingness/acceptance from all stakeholders (Survey 81)
Support for access for mental health professionals	4	First I want to be able to test the programs myself (Survey 93)
Program evaluation	4	Evidence-based programs (Survey 121)
No desire for more BT	3	Not everything needs to go into the direction of digitalization (Survey 25)

[a]n: number of participants with code.

Discussion

Principal Findings

This study explored mental health professionals' perceptions of and experiences with BT. The quantitative findings indicate that participants report little knowledge of BT. Attitude toward BT was somewhat positive, and the acceptance of BT was moderate, comparable to previous literature from German-speaking countries [21] but divergent from a survey conducted with mental health professionals in the Netherlands where perceptions were generally positive [28]. This points to different perceptions of BT depending on the country in question and potential differing experiences with BT in different countries (see also Topooco et al [18] for a survey on attitudes toward digital interventions examined in different European countries). In Switzerland specifically, BT is not routinely implemented yet, and several applications of BT are currently not reimbursed by basic health insurance models. This specific barrier has also been highlighted in an interview study with executive staff and leadership of different Swiss psychiatric institutions, where cost coverage was mentioned as an important aspect [23].

In addition, during recruitment, several professional associations, training institutions, and clinics actively declined to distribute the survey, citing staff shortages, an overload of inquiries, or other individual reasons. These experiences during recruitment may themselves potentially be indicative of broader attitudes toward blended therapy. Specifically, limited time resources or competing institutional priorities might reflect not only organizational constraints but also a lower perceived relevance or priority of BT within some professional contexts. Conversely, the fact that a considerable number of institutions were willing to disseminate the survey may point to growing awareness and openness toward the topic. This recruitment pattern could therefore indirectly mirror varying levels of acceptance or interest in BT among institutions and professionals, a finding that warrants further exploration in future research.

While most participants in our study reported some prior experience with BT, participants rarely used BT in the past 4 weeks. Additionally, the results revealed significant differences in the utilization of various digitally delivered intervention formats for BT, with teletherapy (video) being the most frequently used. Regarding suitability for BT, our study found significant differences between digitally delivered intervention types. Moreover, in our study, BT was deemed suitable for

Mood disorders and Neurotic, stress-related, and Somatoform disorders by most participants (more than 75%, n=152), but suitable for Schizophrenia and delusional disorders or Intellectual disabilities by less than 20% of the participants. This may again in part be related to a lack of knowledge on BT, as studies have shown that digitally delivered interventions can also be successful as add-ons to treatment as usual for patients with schizophrenia-spectrum disorders [29] and that BT can be feasible for severe mental health disorders [30]. Willingness to use BT differed significantly between different treatment points. Descriptively, participants gave the lowest ratings for digitally delivered interventions as a substitute for face-to-face sessions. Willingness to use BT differed significantly across settings, with the lowest acceptance reported for acute inpatient care. This finding contrasts studies conducted on BT in the acute patient setting that show that stakeholders in acute inpatient care consider BT a suitable and relevant treatment option [31].

The qualitative analysis highlighted both perceived advantages and disadvantages of BT. Participants felt that BT can offer benefits, with therapy factors such as flexibility, outsourcing elements, and efficiency being most common. This aligns with the findings from a pilot trial on BT in Swiss outpatient care, where work independent of place and time was mentioned as a positive aspect of BT by therapists [32]. In our survey, patient factors included increased self-efficacy and lowered barriers to therapy. Strengthened therapeutic relationships and mental health professional−related benefits like enhanced capacity and remote work options further highlighted its practicality and appeal. The disadvantages reported by participants included additional effort, concerns about interpersonal interactions such as interruptions for the therapeutic relationship, and challenges with indication.

Aspects concerning the therapeutic relationship were considered both an advantage and a disadvantage of BT by mental health professionals. Interestingly, research shows that a therapeutic relationship can be established in digitally delivered interventions [33-36] and has, for example, been rated higher in BT than in usual care for depression [37]. This highlights a discrepancy between a polarized perception of the therapeutic relationship in BT by mental health professionals and the findings from empirical data on the therapeutic relationship in BT.

Regarding challenges for BT implementation, perceived hurdles included technical issues such as data security alongside direct and indirect costs. For the future, mental health professionals desire cost coverage of BT, accessibility, and easy integration of digitally delivered interventions into therapy. It should be noted that some of the aspects mentioned regarding cost coverage may be very specific to the Swiss context, where digital mental health interventions are currently mostly not included in basic health insurance models for patients.

Future Directions

Nationally representative surveys assessing mental health professionals' perceptions and experiences with BT should be conducted. In addition, it would be of interest to compare patient and mental health professional perspectives of BT using survey-based assessments. Moreover, longitudinal assessments should be used to examine BT perception changes over time. Finally, one future direction that seems particularly clinically relevant is to find effective ways of increasing knowledge on BT among therapy providers. This can be achieved by advancing information on BT in psychotherapy training but also by increasing exposure to digital interventions.

Strengths and Limitations

To the best of our knowledge, this study is the first to investigate the topic of BT in depth among psychotherapists and psychiatrists (in training) in Switzerland. Recruitment strategies were broad (institutions, professional associations, clinics) with the aim of including a broad range of participants. Along with general perceptions of BT, modality-specific information was gained. Moreover, quantitative and qualitative methods were combined to analyze the data. The study also has limitations. First, the survey is not a representative sample of all psychotherapists and psychiatrists in Switzerland. It may have been biased, as only mental health professionals interested in BT filled out the survey. In addition, the distribution of professional experience in our sample was skewed, with more than one third of the participants reporting over 15 years of work experience, while only a small proportion had little or no experience. This uneven representation of experience levels limits the generalizability of our findings. Furthermore, our sample included different groups (eg, professional group or being in training vs not or therapeutic orientations). As shown in our multimedia appendices, some groups differed with regard to, for example, the use of specific digital interventions for BT. Moreover, the findings for a Swiss convenience sample may not translate to the perception of BT in other countries where, for example, attitudes toward digitally delivered intervention are more positive. Third, only a very short definition of BT was provided at the beginning of the survey. Thus, the concepts of BT may have differed widely between participants. While we decided to include the combination of teletherapy (video) and face-to-face sessions in our definition of BT, other studies have taken a different approach. Some equate videotherapy more with face-to-face treatment. Moreover, blended treatment has also been described as the combination of digital intervention and videotherapy [38]. Finally, the reported analyses provide a predominantly descriptive picture of cross-sectional data.

Conclusions

While BT offers an innovative treatment option for patients with mental health disorders, mental health professionals report little knowledge, a somewhat positive attitude, and moderate acceptance. Both advantages and disadvantages of BT as perceived by mental health professionals were detailed in this study. Future implementation may be aided by increasing knowledge on BT for mental health professionals and in the Swiss context specifically by improving cost coverage options.

Acknowledgments

We would like to thank all survey participants for their participation.

Funding

Funding was provided by the Department of Clinical Psychology and Psychotherapy at the University of Bern, Switzerland. Open access funding was provided by the University of Bern.

Data Availability

A de-identified dataset is available from the corresponding author upon reasonable request.

Authors' Contributions

AMK and LLB wrote the initial version of the manuscript. AMK, KG, and EM conducted data analysis. EK set up the online survey. LLB, TB, EM, and KG conceptualized the survey items. LLB conceptualized the study. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey.
[DOCX File, 5660 KB - mental_v13i1e78079_app1.docx]

Multimedia Appendix 2
Research questions and domains.
[DOCX File, 3890 KB - mental_v13i1e78079_app2.docx]

Multimedia Appendix 3
Information on missing data.
[DOCX File, 3694 KB - mental_v13i1e78079_app3.docx]

Multimedia Appendix 4
Supplementary analyses.
[DOCX File, 2234 KB - mental_v13i1e78079_app4.docx]

References

1. Andersson G, Carlbring P, Titov N, Lindefors N. Internet interventions for adults with anxiety and mood disorders: a narrative umbrella review of recent meta-analyses. Can J Psychiatry 2019 Jul;64(7):465-470. [doi: 10.1177/0706743719839381] [Medline: 31096757]
2. Nunes-Zlotkowski KF, Shepherd HL, Beatty L, Butow P, Shaw JM. Blended psychological therapy for the treatment of psychological disorders in adult patients: systematic review and meta-analysis. Interact J Med Res 2024 Oct 29;13:e49660. [doi: 10.2196/49660] [Medline: 39470720]
3. Hedman-Lagerlöf E, Carlbring P, Svärdman F, Riper H, Cuijpers P, Andersson G. Therapist-supported Internet-based cognitive behaviour therapy yields similar effects as face-to-face therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. World Psychiatry 2023 Jun;22(2):305-314. [doi: 10.1002/wps.21088] [Medline: 37159350]
4. Karyotaki E, Riper H, Twisk J, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry 2017 Apr 1;74(4):351-359. [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
5. Karyotaki E, Efthimiou O, Miguel C, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 1;78(4):361-371. [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
6. Berger T, Krieger T, Sude K, Meyer B, Maercker A. Evaluating an e-mental health program ("deprexis") as adjunctive treatment tool in psychotherapy for depression: results of a pragmatic randomized controlled trial. J Affect Disord 2018 Feb;227:455-462. [doi: 10.1016/j.jad.2017.11.021] [Medline: 29154168]

7. Nakao S, Nakagawa A, Oguchi Y, et al. Web-based cognitive behavioral therapy blended with face-to-face sessions for major depression: randomized controlled trial. J Med Internet Res 2018 Sep 21;20(9):e10743. [doi: 10.2196/10743] [Medline: 30249583]
8. Bielinski LL, Trimpop L, Berger T. All in the mix? Blended psychotherapy as an example of digitalization in psychotherapy. Psychotherapeut 2021;66(5):447-454. [doi: 10.1007/s00278-021-00524-3] [Medline: 34257478]
9. Erbe D, Eichert HC, Riper H, Ebert DD. Blending face-to-face and internet-based interventions for the treatment of mental disorders in adults: systematic review. J Med Internet Res 2017 Sep 15;19(9):e306. [doi: 10.2196/jmir.6588] [Medline: 28916506]
10. Zwerenz R, Becker J, Knickenberg RJ, Siepmann M, Hagen K, Beutel ME. Online self-help as an add-on to inpatient psychotherapy: efficacy of a new blended treatment approach. Psychother Psychosom 2017;86(6):341-350. [doi: 10.1159/000481177] [Medline: 29131090]
11. Kemmeren LL, van Schaik A, Draisma S, Kleiboer A, Riper H, Smit JH. Effectiveness of blended cognitive behavioral therapy versus treatment as usual for depression in routine specialized mental healthcare: E-COMPARED trial in the Netherlands. Cogn Ther Res 2023 Jun;47(3):386-398. [doi: 10.1007/s10608-023-10363-y]
12. Schaeuffele C, Mutak A, Behr S, et al. Increasing the effectiveness of psychotherapy in routine care through transdiagnostic online modules? Randomized controlled trial investigating blended care. PsyArXiv. Preprint posted online on 2024. [doi: 10.31234/osf.io/972z5]
13. Schuster R, Pokorny R, Berger T, Topooco N, Laireiter AR. The advantages and disadvantages of online and blended therapy: survey study amongst licensed psychotherapists in Austria. J Med Internet Res 2018 Dec 18;20(12):e11007. [doi: 10.2196/11007] [Medline: 30563817]
14. Vis C, Mol M, Kleiboer A, et al. Improving implementation of eMental health for mood disorders in routine practice: systematic review of barriers and facilitating factors. JMIR Ment Health 2018 Mar 16;5(1):e20. [doi: 10.2196/mental.9769] [Medline: 29549072]
15. Vis C, Kleiboer A, Prior R, et al. Implementing and up-scaling evidence-based eMental health in Europe: the study protocol for the MasterMind project. Internet Interv 2015 Nov;2(4):399-409. [doi: 10.1016/j.invent.2015.10.002]
16. Békés V, Aafjes-van Doorn K, Luo X, Prout TA, Hoffman L. Psychotherapists' challenges with online therapy during COVID-19: concerns about connectedness predict therapists' negative view of online therapy and its perceived efficacy over time. Front Psychol 2021;12:705699. [doi: 10.3389/fpsyg.2021.705699] [Medline: 34367030]
17. Staeck R, Drüge M, Albisser S, Watzke B. Acceptance of E-mental health interventions and its determinants among psychotherapists-in-training during the first phase of COVID-19. Internet Interv 2022 Sep;29:100555. [doi: 10.1016/j.invent.2022.100555] [Medline: 35789691]
18. Topooco N, Riper H, Araya R, et al. Attitudes towards digital treatment for depression: a European stakeholder survey. Internet Interv 2017 Jun;8(1–9):1-9. [doi: 10.1016/j.invent.2017.01.001] [Medline: 30135823]
19. Schuster R, Topooco N, Keller A, Radvogin E, Laireiter AR. Advantages and disadvantages of online and blended therapy: replication and extension of findings on psychotherapists' appraisals. Internet Interv 2020 Sep;21:100326. [doi: 10.1016/j.invent.2020.100326] [Medline: 32477885]
20. van der Vaart R, Witting M, Riper H, Kooistra L, Bohlmeijer ET, van Gemert-Pijnen LJ. Blending online therapy into regular face-to-face therapy for depression: content, ratio and preconditions according to patients and therapists using a Delphi study. BMC Psychiatry 2014 Dec 14;14(355):355. [doi: 10.1186/s12888-014-0355-z] [Medline: 25496393]
21. Mittmann G, Steiner-Hofbauer V, Schrank B. Attitudes of the general population and mental health practitioners towards blended therapy in Austria. Wien Klin Wochenschr 2025 Feb;137(3-4):118-125. [doi: 10.1007/s00508-024-02391-9] [Medline: 39037450]
22. Braun P, Drüge M, Hennemann S, Nitsch FJ, Staeck R, Apolinário-Hagen J. Acceptance of e-mental health services for different application purposes among psychotherapists in clinical training in Germany and Switzerland: secondary analysis of a cross-sectional survey. Front Digit Health 2022;4:840869. [doi: 10.3389/fdgth.2022.840869] [Medline: 35295621]
23. Best M, Bielinski LL, Berger T. Implementierung digitaler Interventionen in psychiatrischen Kliniken der Schweiz. Psychiatrie + Neurologie 2022:9-12 [FREE Full text]
24. Qualtrics XM platform [computer software]. Qualtrics. 2023. URL: https://www.qualtrics.com [accessed 2025-12-30]
25. ICD-10-GM 2024: internationale statistische klassifikation der krankheiten, 10 revision, German modification version 2024. ICD-Code.de.: Bundesinstitut für arzneimittel und medizinprodukte (BFARM) URL: https://www.icd-code.de [accessed 2025-04-11]
26. Mayring P. Qualitative Inhaltsanalyse: Grundlagen Und Techniken: Beltz Verlagsgruppe; 2022. URL: https://www.beltz.de/fileadmin/beltz/leseproben/978-3-407-25898-4.pdf [accessed 2025-12-30]
27. Schweizerisches institut für ärztliche weiter- und fortbildung (SIWF) SIWF – Swiss Institute for Postgraduate and Continuing Medical Education. SIWF.ch. 2023. URL: https://www.siwf.ch [accessed 2023-07-23]
28. Dijksman I, Dinant GJ, Spigt M. The perception and needs of psychologists toward blended care. Telemed J E Health 2017 Dec;23(12):983-995. [doi: 10.1089/tmj.2017.0031] [Medline: 28556693]
29. Westermann S, Rüegg N, Lüdtke T, Moritz S, Berger T. Internet-based self-help for psychosis: findings from a randomized controlled trial. J Consult Clin Psychol 2020 Oct;88(10):937-950. [doi: 10.1037/ccp0000602] [Medline: 32790453]

30. Ehrt-Schäfer Y, Rusmir M, Vetter J, Seifritz E, Müller M, Kleim B. Feasibility, adherence, and effectiveness of blended psychotherapy for severe mental illnesses: scoping review. JMIR Ment Health 2023 Dec 26;10:e43882. [doi: 10.2196/43882] [Medline: 38147373]
31. Bielinski LL, Wälchli G, Lange A, et al. A qualitative analysis of healthcare professionals' experiences with an internet-based emotion regulation intervention added to acute psychiatric inpatient care. BMC Psychiatry 2024 Dec 27;24(1):955. [doi: 10.1186/s12888-024-06365-z] [Medline: 39731056]
32. Bielinski LL, Bur OT, Wälchli G, et al. Two sides of the same coin? Patient and therapist experiences with a transdiagnostic blended intervention focusing on emotion regulation. Internet Interv 2022 Dec;30:100586. [doi: 10.1016/j.invent.2022.100586] [Medline: 36386404]
33. Berger T. The therapeutic alliance in internet interventions: a narrative review and suggestions for future research. Psychother Res 2017 Sep;27(5):511-524. [doi: 10.1080/10503307.2015.1119908] [Medline: 26732852]
34. Flückiger C, Del Re AC, Wampold BE, Horvath AO. The alliance in adult psychotherapy: a meta-analytic synthesis. Psychotherapy (Chic) 2018 Dec;55(4):316-340. [doi: 10.1037/pst0000172] [Medline: 29792475]
35. Pihlaja S, Stenberg JH, Joutsenniemi K, Mehik H, Ritola V, Joffe G. Therapeutic alliance in guided internet therapy programs for depression and anxiety disorders—a systematic review. Internet Interv 2018 Mar;11:1-10. [doi: 10.1016/j.invent.2017.11.005] [Medline: 30135754]
36. Probst GH, Berger T, Flückiger C. The alliance-outcome relation in internet-based interventions for psychological disorders: a correlational meta-analysis. Verhaltenstherapie 2022;32(Suppl. 1):135-146. [doi: 10.1159/000503432]
37. Doukani A, Quartagno M, Sera F, et al. Comparison of the working alliance in blended cognitive behavioral therapy and treatment as usual for depression in Europe: secondary data analysis of the E-COMPARED randomized controlled trial. J Med Internet Res 2024 May 31;26:e47515. [doi: 10.2196/47515] [Medline: 38819882]
38. Etzelmueller A, Radkovsky A, Hannig W, Berking M, Ebert DD. Patient's experience with blended video- and internet based cognitive behavioural therapy service in routine care. Internet Interv 2018 Jun;12:165-175. [doi: 10.1016/j.invent.2018.01.003] [Medline: 30135780]

Abbreviations

BT: blended therapy
***ICD-10*:** *International Statistical Classification of Diseases, Tenth Revision*
MD: mean difference

Edited by J Torous; submitted 27.May.2025; peer-reviewed by C Schaeuffele, I Titzler; revised version received 07.Nov.2025; accepted 24.Nov.2025; published 06.Jan.2026.

Please cite as:
Kneubühler AM, von Känel E, Grgic K, Munkovic E, Berger T, Bielinski LL
Blended Therapy From the Perspective of Mental Health Professionals in Routine Mental Health Care: Mixed Methods Analysis of Cross-Sectional Survey Data
JMIR Ment Health 2026;13:e78079
URL: https://mental.jmir.org/2026/1/e78079
doi:10.2196/78079

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org