Effectiveness of a fully automated mobile therapeutic versus general chatbot in reducing depression and anxiety and improving well-being: A pilot study

Barbora Kutá, Lukáš Novák, Radka Žídková, Jana Fürstová, Klára Maliňáková, Andrea De Winter, Vít Hušek

Submitted to: JMIR Mental Health
on: August 27, 2025

Table of Contents

Effectiveness of a fully automated mobile therapeutic versus general chatbot in reducing depression and anxiety and improving well-being: A pilot study

Barbora Kutá[1]; Lukáš Novák[1]; Radka Žídková[1]; Jana Fürstová[1]; Klára Mali?áková[1]; Andrea De Winter[2]; Vít Hušek[1]

[1] Palacký University Olomouc Olomouc CZ
[2] University Medical Center Groningen Groningen NL

Corresponding Author:
Barbora Kutá

Palacký University Olomouc
K?ížkovského 511/8
Olomouc
Olomouc
CZ

Abstract

Background: Given the increasing prevalence of depression and anxiety disorders and enduring barriers to care, there is a critical need for alternative treatment options. Generative AI chatbots show promise for increasing access to mental health care, though more direct research is needed to establish their efficacy.

Objective: This pilot study tested the efficacy of a generative mental health chatbot rooted in solution-focused therapy compared to the general-purpose ChatGPT and an assessment only control (AOC) group on depression, anxiety, and well-being.

Methods: A total of 185 online recruited English-speaking adults were randomly assigned to one of three groups: AI therapy, ChatGPT, or AOC. Of these, 147 eligible participants filled out a pretreatment assessment. Over a three-week period, the AI therapy group (n=44) was instructed to complete three structured, fully automated app-based sessions per week (9 total), while the ChatGPT group (n=60) was instructed to engage in 9 unstructured conversations with ChatGPT (GPT-4o-based models). The control group (n=43) received no intervention. In AI therapy group, 39% completed all sessions, and 62% of those in the ChatGPT group. Primary outcomes measures, self-assessed online at baseline and post-intervention, included PHQ-9, ODSIS (depression), GAD-7 (anxiety), and WHO-5 (well-being). Linear mixed-effects models were employed for data analysis.

Results: Compared to AOC, both the AI therapy group (d = -0.47, P = 0.012) and the ChatGPT group (d = -0.44, P = 0.023) demonstrated significant reductions in depression scores measured by PHQ-9. The AI therapy group showed non-significant reductions in anxiety (d = -0.37, p =0.106), ODSIS depression scores (d = -0.25, P = 0.220), and an increase in well-being (d = 0.12; P = 0.525) compared to AOC. Similarly, a non-significant reduction in anxiety (d = -0.27, P = 0.220), ODSIS depression scores (d = -0.12, P = 0.530), and an increase in well-being (d = 0.20, P = 0.285) were observed in the ChatGPT group compared to AOC. The AI therapy group did not significantly outperform the ChatGPT group on any outcomes (PHQ-9: b = -0.19, d = 0.03, P = 0.874; GAD-7: b = -0.57, d = -0.11, P = 0.621; ODSIS: b = -0.59, d = -0.13, P = 0.498; WHO: b = -0.38, d = -0.07, P = 0.691).

Conclusions: Both the structured generative AI chatbot and ChatGPT showed a significant reduction in depression scores compared to the control group. No significant effects were observed across other outcomes, although descriptive trends indicated improvements in anxiety. While the AI therapy group showed descriptively better outcomes for depression and anxiety, differences between groups were not significant. A larger sample and longer intervention may be needed for the emerging trends to yield clinically meaningful effect sizes. Clinical Trial: OSF Registries https://osf.io/r76ef

(JMIR Preprints 27/08/2025:82642)
DOI: https://doi.org/10.2196/preprints.82642

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Title:

Effectiveness of a fully automated mobile therapeutic versus general chatbot in reducing depression and anxiety and improving well-being: A pilot study

Authors:
Barbora Kuta: 0009-0003-0494-3890, Palacký University Olomouc
Lukas Novak: 0000-0002-7582-2098, Palacký University Olomouc
Radka Zidkova: 0000-0003-3144-2437, Palacký University Olomouc
Jana Furstova: 0000-0003-4057-2540, Palacký University Olomouc
Klára Maliňáková: 0000-0001-6939-1204, Palacký University Olomouc
Andrea De Winter: 0000-0002-9380-4889, University of Groningen
Vit Husek: 0000-0001-6989-2383, Palacký University Olomouc

Abstract

Background:
Given the increasing prevalence of depression and anxiety disorders and enduring barriers to care, there is a critical need for alternative treatment options. Generative AI chatbots show promise for increasing access to mental health care, though more direct research is needed to establish their efficacy.

Objective:
This pilot study tested the efficacy of a generative mental health chatbot rooted in solution-focused therapy compared to the general-purpose ChatGPT and an assessment only control (AOC) group on depression, anxiety, and well-being.

Methods:
A total of 185 online recruited English-speaking adults were randomly assigned to one of three groups: AI therapy, ChatGPT, or AOC. Of these, 147 eligible participants filled out a pretreatment assessment. Over a three-week period, the AI therapy group (n=44) was instructed to complete three structured, fully automated app-based sessions per week (9 total), while the ChatGPT group (n=60) was instructed to engage in 9 unstructured conversations with ChatGPT (GPT-4o-based models). The control group (n=43) received no intervention. In AI therapy group, 39% completed all sessions, and 62% of those in the ChatGPT group. Primary outcomes measures, self-assessed online at baseline and post-intervention, included PHQ-9, ODSIS (depression), GAD-7 (anxiety), and WHO-5 (well-being). Linear mixed-effects models were employed for data analysis.

Results:
Compared to AOC, both the AI therapy group (d = -0.47, *P* = 0.012) and the ChatGPT group (d = -0.44, *P* = 0.023) demonstrated significant reductions in depression scores measured by PHQ-9. The AI therapy group showed non-significant reductions in anxiety (d = -0.37, p =0.106), ODSIS depression scores (d = -0.25, *P* = 0.220), and an increase in well-being (d = 0.12; *P* = 0.525) compared to AOC. Similarly, a non-significant reduction in anxiety (d = -0.27, *P* = 0.220), ODSIS depression scores (d = -0.12, *P* = 0.530), and an increase in well-being (d = 0.20, *P* = 0.285) were observed in the ChatGPT group compared to AOC. The AI therapy group did not significantly outperform the ChatGPT group on any outcomes (PHQ-9: b = -0.19, d = 0.03, *P* = 0.874; GAD-7: b = -0.57, d = -0.11, *P* = 0.621; ODSIS: b = -0.59, d = -0.13, *P* = 0.498; WHO: b = -0.38, d = -0.07, *P* = 0.691).

Conclusions:
Both the structured generative AI chatbot and ChatGPT showed a significant reduction in depression scores compared to the control group. No significant effects were observed across other outcomes, although descriptive trends indicated improvements in anxiety. While the AI therapy group showed descriptively better outcomes for depression and anxiety, differences between groups were not significant. A larger sample and longer intervention may be needed for the emerging trends to yield clinically meaningful effect sizes.

Study Registration:
OSF Registries https://osf.io/r76ef

Keywords: depression; anxiety; well-being; conversational agents; digital intervention; randomized controlled trial; chatbot; solution-focused therapy

Introduction

One in every seven people meet criteria for a mental health disorder in 2021, with approximately 229 million people globally dealing with depression and 359 million with anxiety disorders [1]. Mental health illnesses have moved from the 9th to the 6th leading causes of disability-adjusted life years from 1990 to 2021 [2]. This shift highlights the growing societal burden of mental illness. Alongside the escalating prevalence in mental illness, the treatment gap is widening with estimates that fewer than 50% of adults with mental illness receive any form of mental health treatment, even in high-income countries [3].

It is estimated that 58% of people with clinical-level mental health issues don't seek any professional help [4], as access to care remains a salient barrier. Due to the lack of treatment services, only 23% of people affected by depression receive minimally adequate treatment according to current research standards in high-income countries, and even fewer (8%) receive such treatment in low and lower-middle-income countries [5]. Additionally, the waiting times for mental health services are prohibitive, averaging longer than 3 months [6,7]. Another barrier is financial affordability, though its impact has slightly decreased in recent years [8]. Moreover, in addition to these systemic factors, individual factors may also explain reticence to seek care. People affected by mental illness report a low perceived need for treatment, a desire to handle problems independently [9,10], or feeling too busy to pursue treatment [8]. In light of these findings, it is critical to establish treatment options that are affordable, easily accessible, and, if possible, short-term and effective.

Digital mental health intervention (DMHIs) tools are promising supplements and/or alternatives to traditional treatment that can enhance mental health care accessibility [11]. Digital modalities are gaining traction including online psychotherapy and programs powered by virtual reality [12]. Anonymity is a key advantage of DMHIs; people who believe they are talking to a computer, relative to a human operator, are more willing to disclose or express sadness and have less fear of being evaluated [13]. Subsequently, DMHI tools have growing support for treating mental health challenges, including depression [14] and anxiety [15]. Moreover, when combined with traditional therapy, mental health apps that support patients between sessions can improve the effectiveness of usual treatment for both depressive and anxiety disorders [16].

With the rise of artificial intelligence (AI), particularly through the development of Large Language Models (LLMs), new possibilities have emerged to mediate the primary psychotherapeutic tool, conversation. Conversational agents (CAs) or chatbots represent a promising, accessible, and affordable mental health tool that could automate some therapeutic procedures when the demand for professionals exceeds available capacity [17]. This leverages one of the primary benefits of DMHIs, which is a personalization of the treatment, especially using machine learning [18], potentially leading to greater positive outcomes and lower dropout rates [11,19]. AI has reached a point where it is challenging to differentiate real conversations from those with CAs. When therapists were asked to distinguish between transcripts of interactions with human therapists and those with AI chatbots, they were correct only 53.9% of the time, performing no better than random guessing [20].

As early as 2019, at least 41 mental health chatbots were on the market, most of them claiming to provide therapy [21]**.** Yet, many of them had not been reviewed by professionals, placing them in a regulatory 'grey area' and raising questions of safety that have been discussed only by a few studies so far [22]. For professionals to accept and build trust in the new technology, it is crucial to have evidence-based tools [17,23], but rigorous research in this area has lagged behind AI's rapid development. The advancement of generative AI (GenAI) in the 2020s has significantly bolstered the capabilities of mental health CAs [24]. Although meta-analyses comparing randomized controlled trials (RCTs) on the effectiveness of CAs already exist, they include studies of different types of chatbots in terms of function (retrieval-based, rule-based, generative) [25–28]. A systematic review and meta-analysis from 2024 showed a significant effect of CAs on depression, but did not discuss the type of chatbot or AI model [25]. Another review and meta-analysis from 2023 showed similar positive results, but the majority of included studies examined retrieval-based CAs [27]. Even the most recent meta-analysis from 2025 on young people included only 3 studies examining GenAI CAs [28]. Therefore, as there are still only a few studies exploring the effect of GenAI chatbots on mental health, more evidence is needed to examine their effectiveness.

Generative AI chatbots could help fill the treatment gap by supporting individuals who are waiting for therapy, who can't afford therapy, and those at low or medium acuity levels who do not need intensive treatment or would not otherwise engage in therapy. The present study entails an early-stage pilot trial testing the merits of a GenAI therapy chatbot. As described below, the AI therapy chatbot tested in this pilot was trained to deliver Solution-Focused Brief Therapy (SFBT), a therapeutic approach selected for its strong fit with chatbot delivery. Specifically, SFBT offers a structured conversational format and emphasizes brief, goal-oriented interactions. These features make it especially well-suited for an initial pilot focused on feasibility and short-term clinical outcomes. In this pilot RCT, the AI therapy chatbot was compared to both an untrained ChatGPT-4o-based chatbot and an assessment only control (AOC) group. Given the non-clinical and non-treatment-seeking sample, the study focused on short-term outcomes including symptoms of depression and anxiety, as well as well-being indices.

Methods

Study design and participants

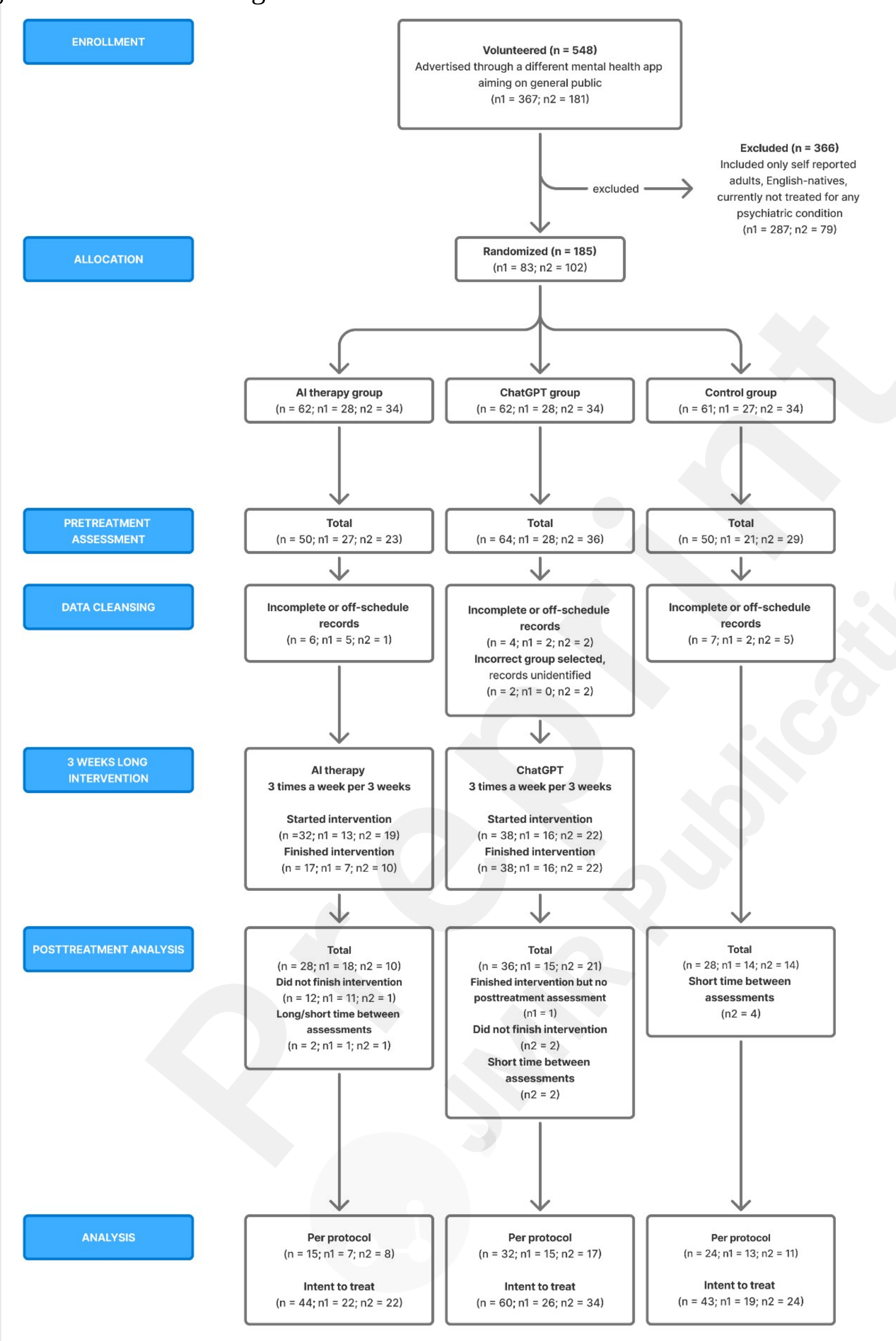
The preregistered study entailed a pilot randomized controlled trial with participants randomly allocated in a 1:1:1 ratio across three groups: (1) The AI Therapy group who engaged in AI-assisted therapy sessions via the ChatMind app with a requirement of three sessions per week for three weeks; (2) The ChatGPT group who completed equivalent sessions with a general chatbot in the ChatGPT mobile app; and (3) The AOC group who received no intervention. Mental health indices were measured before and after the three-week intervention protocols using an online self-assessing questionnaire hosted on the OQS platform.

The sample size was estimated based on a previous RCT [29] evaluating a 3-week solution-focused brief, human-delivered therapy intervention, as the AI therapy shares similar characteristics. Using the effect sizes from this study, a range of anticipated mean changes were estimated and a simulation-based power analysis was conducted. Across different sample sizes, specifying a mixed-effects model accounting for repeated measures design and group differences, power simulations indicated that 16 participants per group would achieve ≥80% power to detect a very large effect size ($d = 1.24$). However, given the aims of this pilot-stage study, statistical power to detect a smaller effect was not the priority. More details on this analysis, including a power analysis report, are described in the project pre-registration.

Participants were recruited using a convenience sampling method through online advertisements (e.g., Instagram), e-mails, and push notifications in a partnered mental health app (i.e., VOS, a daily mood tracker and self-care tool). Inclusion criteria were minimal; participants were eligible if they were native English-speaking, aged 18+, were not being treated for any psychiatric condition, and had never used the ChatMind AI chatbot. Participation was voluntary with the only incentive being trial access to the ChatMind AI chatbot (i.e., participants in all conditions got access after the three-week study period). Informed consent was obtained from all individuals before the baseline assessment. The study design was approved by the Ethics Committee at Olomouc University Social Health Institute - OUSHI - (Approval No. 2024/09).

Participants were recruited in two waves. First, 83 participants were enrolled in November/December 2024, but only 76 completed the baseline survey (92%), of which 9 records were deleted for completing the questionnaire outside the time schedule or not finishing the questionnaire (12%), resulting in 67 randomized participants in the first wave. To ensure a sufficient sample size for feasibility testing and pilot-stage effect size estimates, an additional recruitment wave took place in February 2025. Of the 102 participants enrolled in this wave, 88 completed the baseline survey (86%), of which 8 were incomplete or off-schedule records, resulting in 80 participants in the second wave and 147 participants in total. Consistent with the aims of a pilot RCT, statistical power was not prioritized; instead, the sample size was selected to enable estimation of preliminary effect sizes to provide early-stage evidence and foundational support for forthcoming large-scale trials. The CONSORT diagram chart in Firuge 1 illustrates entire process of respondent enrollment, allocation into groups, assessment, and intervention.

Figure 1. CONSORT diagram chart



The final analytic sample comprised 85 participants from the U.S. (58%), 21 from Canada (14%), and 41 from other predominantly English-speaking countries (28%) including the U.K., Ireland, and Australia. Participants ranged in age from 20 to 74 years old (M_{age}=38.4, SD=10.8), and 73% were women.

Measures

Anxiety Symptoms

Anxiety was measured using the Generalized Anxiety Disorder Scale (GAD-7), a 7-item scale assessing anxiety symptoms over the past two weeks, rated on a 4-point scale (0 = *"Not at all"* to 3 = *"Nearly every day"*). Total scores range from 0 to 21, with higher scores indicating greater anxiety severity [30]. Cronbach's alpha was 0.89 using both baseline and post intervention data.

Depressive Symptoms

Depressive symptoms were evaluated with two scales, the Patient Health Questionnaire (PHQ-9) [31] and Overall Depression Severity and Impairment Scale (ODSIS) [32]. The PHQ-9 assesses the severity of depressive symptoms over the past two weeks through nine items based on the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) criteria for depression. Responses indicate the frequency of symptoms and are rated on a 4-point scale (0 = *"Not at all"* to 3 = *"Nearly every day"*). Total scores range from 0 to 27, with higher scores indicating more severe symptoms [31]. Cronbach's alpha was 0.87 using both pre and post intervention data. The ODSIS measures the severity and functional impairment of depression. Participants responded to items indicating symptom frequency on a 5-point scale (0 = *"Not at all"* to 4 = *"All the time"*). Total scores range from 0 to 20, whereby higher scores indicate greater impairment [32]. Based on both baseline and post intervention data, Cronbach's alpha was 0.94.

Well-being

The Well-being Index (WHO-5) was utilized to assess subjective well-being. Participants indicated how often they had experienced any of the manifestations of well-being in the past two weeks on a 6-point scale (0 = *"At no time"* to 5 = *"All of the time"*), with total scores ranging from 0 to 25 and higher scores indicating greater well-being [33,34]. The Cronbach's alpha of this tool was 0.90 using the baseline and post intervention data.

For each scale, we used sum scores treated as continuous variables.

Intervention

AI therapy

Participants in the AI therapy group were given access to the ChatMind app that provided AI therapy rooted in solution-focused brief therapy (SFBT) principles via voice and text messaging. The app consists of 2 types of session lengths, roughly 10 and 30 minutes, but the final length depended on the flow of the conversation. Participants were instructed to do one short and one long session, and a third session of their choice per week.

During each session, the AI therapy chatbot first identified the participant's problem and expectations. Then it guided the participant to find a solution or a small step they can take to improve their situation. Conversations operate on generative AI, modified through prompt engineering and specialized architecture. Participants could choose to engage in oral or text communication.

ChatGPT Group

Participants in this group were asked to download the ChatGPT mobile app, which allows both text and oral communication. From November 2024 to February 2025, free-tier ChatGPT users initially used GPT-4o model, and after reaching their message limit, the system automatically switched them to the lighter GPT-4o-mini model. Therefore, for most conversations the 4o model was used, and 4o-mini for the remainder. Participants were instructed to interact with ChatGPT 3 times a week for at least 10 minutes, as if they were talking to an AI therapy chatbot about whatever was bothering them.

Assessment Only Control Group

Participants in the AOC group completed baseline and follow-up assessments three weeks later, but did not receive any intervention components. AOC participants were instructed not to use any chatbot for psychological intervention during the three-week study period.

Statistical analysis

Group differences in demographic variables and daseline differences in outcome measures, were estimated using chi-square tests for categorical variables and non-parametric analysis of variance (the Kruskal-Wallis test) for age. A chi-squared test was also performed to assess for differential attrition across the study groups. This test revealed a strong trend suggesting that dropout rates were dependent on group assignment ($\chi^2 = 5.14$, $P = 0.082$). This signal of differential attrition, driven by a substantially higher dropout rate in the AI Therapy group, raises concerns about the validity of a per-protocol analysis, as this approach becomes susceptible to selection bias that can compromise the initial randomization. Thus, our primary analysis followed the intention-to-treat (ITT) principle to provide an unbiased estimate of the intervention's effectiveness. A secondary per-protocol analysis (PPA), consistent with one of the options outlined in our preregistration, was also conducted to explore the efficacy of the intervention specifically among participants who completed the study. The results of the PPA are presented in Supplementary Analysis 1 online.

To evaluate intervention effects, linear mixed-effects models were employed. Separate models were fit for each outcome variable: anxiety (GAD-7), depression (ODSIS, PHQ-9), and mental well-being (WHO-5). For the GAD-7, ODSIS, and WHO-5 outcomes, models were fit, with parameters estimated using restricted maximum likelihood (REML). For the PHQ-9 outcome, initial diagnostic checks revealed significant heteroscedasticity. To address this, a mixed-effects model was refit with the non-constant variance explicitly modelled.

Each model included fixed effects for time, with measurements taken at baseline and after the three-week study period (i.e., pre- vs. post-intervention), experimental group (AI therapy, ChatGPT, AOC), and the time$\times$group interaction representing the differential change over time by group. Random intercepts for participants accounted for individual differences in baseline scores and changes over time. Initially, AOC was coded as the reference group to derive contrasts between AOC and the two active treatment conditions, then we re-leveled the condition variable coding to directly contrast AI therapy and ChatGPT conditions via planned comparisons. To formally assess the mechanism of missing data, we performed an omnibus test for data being Missing Completely at Random (MCAR). In the MCAR test, the assumptions of multivariate normality and homoscedasticity required for the standard parametric Little's test were violated (Hawkins test, $p < .001$). Therefore, we relied on the robust non-parametric alternative. The result of this test was not statistically significant ($P = 0.061$), providing insufficient evidence from this omnibus test to reject the MCAR null hypothesis. To examine predictors of intervention completion (fidelity) and retention (completing the follow-up survey), we performed a series of logistic regression analyses. In these analyses, independent variables were outcome measures (i.e., PHQ-9, ODSIS, WHO-5), age, gender and attitude towards AI.

Effect sizes were reported as unstandardized regression coefficients (b) and Cohen's d. Statistical significance was set at $p<0.05$. All statistical analyses were conducted using R software, version 4.3.0 (R Core Team, 2023) within the RStudio environment, version 2024.04.2.

Results

Sample demographics and baseline characteristics for the intention-to-treat sample are provided in Table 1. Baseline comparisons showed no significant differences across the three groups for age or any of the clinical outcome measures (GAD-7, ODSIS, PHQ-9, WHO-5), with all p-values > .05. However, a chi-square test revealed a significant difference in the distribution of education levels across the groups ($\chi^2 = 16.07$, $P = 0.014$). The demographic characteristics of the sample are summarized in Table 1.

Table 1. Sample characteristics by assigned group

Variable	AI Therapy (N=44)	ChatGPT (N=60)	Control (N=43)	Overall (N=147)
Age				
Mean (SD)	37 (9)	39 (10)	39 (13)	38 (11)
Median [Min, Max]	36 [23, 71]	37 [20, 74]	38 [22, 73]	37 [20, 74]
Gender: N (%)				
Female	28 (64)	42 (70)	37 (86)	107 (73)
Male	16 (36)	18 (30)	6 (14)	40 (27)
Country: N (%)				
USA	25 (57%)	34 (57%)	26 (60%)	85 (58%)
Canada	5 (11%)	10 (17%)	6 (14%)	21 (14%)
Other	14 (32%)	16 (27%)	11 (26%)	41 (28%)
Education: N (%)				
High school or less	6 (14%)	5 (8.3%)	13 (30%)	24 (16%)
Higher vocational	11 (25%)	12 (20%)	3 (7.0%)	26 (18%)
Bachelor degree	13 (30%)	21 (35%)	19 (44%)	53 (36%)
Master's or PhD	14 (32%)	22 (37%)	8 (19%)	44 (30%)
Economical status: N (%)				
Not currently working	15 (34%)	17 (28%)	13 (30%)	45 (31%)
Employed	25 (57%)	31 (52%)	24 (56%)	80 (54%)
Self-employed	4 (9.1%)	12 (20%)	6 (14%)	22 (15%)
Recruitment phase: N (%)				
First	22 (50%)	26 (43%)	19 (44%)	67 (46%)
Second	22 (50%)	34 (57%)	24 (56%)	80 (54%)
Drop out participants: N (%)				
No	15 (34%)	32 (53%)	24 (56%)	71 (48%)
Yes	29 (66%)	28 (47%)	19 (44%)	76 (52%)

Feasibility and Engagement

Full fidelity to the intervention protocol was relatively low in both treatment conditions; 17 (39%) of those in the AI therapy condition completed all intervention sessions, and 38 (62%) of those in the ChatGPT condition completed all sessions. Logistic regression revealed that none of the baseline clinical measures were significant predictors (GAD-7: odds ratio (OR) = 0.97, P = 0.456; ODSIS: OR = 0.95, P = 0.234; PHQ-9: OR = 0.98, P = 0.641). However, older age was significantly associated with a lower likelihood of completing the intervention (OR = 0.95, P = 0.031).

Study retention, in terms of completing the follow-up survey, was 28 (64%) in the AI Therapy conditions, 36 (60%) in the ChatGPT condition, and 28 (65%) in the AOC condition. Logistic regression revealed that baseline symptom severity was not a significant predictor, although higher baseline ODSIS scores showed a trend towards predicting a lower likelihood of retention (OR = 0.92, P = 0.051). In these models, older age again emerged as a significant predictor of lower retention (OR = 0.95, P = 0.003). Furthermore, a more positive baseline attitude towards AI significantly predicted a higher likelihood of retention (OR = 1.48, P = 0.048).

Descriptive Outcome Statistics

Descriptive statistics for anxiety, depression, and mental well-being scores at baseline and after the three-week intervention period, stratified by the experimental groups, are presented in Appendix 1. Kruskal-Wallis tests showed no significant group differences in mental health indices at baseline (p-value range: 0.43 - 0.94). Although not a major feature of the trial, baseline assessment of attitudes towards AI-based therapy was very favorable (M = 3.05 on a 4-point scale, where 1 = Skeptical/Negative and 4 = Enthusiastic).

The largest decrease in mean scores was observed in the AI therapy group, particularly in anxiety symptoms measured by the GAD-7, which dropped by 1.02 points, and in depressive symptoms, as reflected in both the ODSIS (a decrease of 0.69 points) and the PHQ-9 (1.39 points). Another notable change was observed in the ChatGPT group, which showed a reduction of 1.8 points in PHQ-9 scores.

Table 2 presents the intention-to-treat intervention effect estimates, with unstandardized regression coefficients (b) for the interaction terms representing the estimated difference in change between groups from pre to post-intervention. Table 3 presents the corresponding effect sizes (Cohen’s d) for these interaction effects, along with their 95% confidence intervals.

Relative to AOC, the AI therapy group exhibited a 1.98-point greater reduction in anxiety symptoms (d = -0.37) and a 1.13-point greater reduction in ODSIS depression scores (d = -0.25), though neither of these effects was statistically significant. However, the AI therapy group showed a statistically significant, 2.67-point greater reduction in PHQ depressive symptoms (d = -0.47, P = 0.012). Similarly, the ChatGPT condition showed a significant 2.47-point greater reduction in PHQ depressive symptoms (d = -0.44, P = 0.023), but non-significant improvements in anxiety symptoms (d = -0.27) and ODSIS depression scores (d = -0.12). Neither group showed a significant change in well-being relative to control. As shown in Supplementary Analysis 1, these results were consistent with the per-protocol results.

Table 2. Results of mixed-effects models of temporal changes in anxiety (GAD-7), depression (ODSIS, PHQ-9), and mental well-being (WHO-5) scores across study groups.

	Anxiety (GAD-7[a])		Depression (ODSIS[b])		Depression (PHQ-9[c])		Mental Well-being (WHO-5[d])	
Effect	b[e] (95% CI[f])	*P*-value	b (95% CI)	*P*-value	b (95% CI)	*P*-value	b (95% CI)	*P*-value
Fixed effects								
Intercept	13.27 (8.38, 18.16)	<0.001	11.55 (7.99, 15.11)	<0.001	24.00 (19.69, 28.30)	<0.001	9.99 (5.75, 14.23)	<0.001
Time	0.38 (-1.32, 2.08)	0.660	0.35 (-0.93, 1.64)	0.585	0.64 (-0.60, 1.89)	0.313	-0.25 (-1.66, 1.16)	0.725
Experimental Group								
Control	*Reference*							
AI Therapy	-0.90 (-3.17, 1.37)	0.433	-0.22 (-2.27, 1.84)	0.835	-0.98 (-3.23, 1.26)	0.390	0.84 (-1.60, 3.27)	0.499
ChatGPT	0.23 (-1.89, 2.36)	0.829	0.13 (-1.79, 2.06)	0.891	0.28 (-2.01, 2.56)	0.813	-0.49 (-2.77, 1.79)	0.672
Interaction effects								
Control Group × Time	*Reference*							
AI Therapy Group × Time	-1.98 (-4.39, 0.43)	0.106	-1.13 (-2.94, 0.68)	0.220	-2.67 (-4.74, -0.60)	0.012	0.64 (-1.35, 2.64)	0.525
ChatGPT Group × Time	-1.41 (-3.68, 0.85)	0.220	-0.54 (-2.25, 1.17)	0.530	-2.47 (-4.61, -0.34)	0.023	1.02 (-0.86, 2.90)	0.285
Planned Comparisons								
ChatGPT Group × Time	*Reference*							
AI Therapy Group × Time	-0.57 (-2.83, 1.70)	0.621	-0.59 (-2.29, 1.12)	0.498	-0.19 (-2.59, 2.20)	0.874	-0.38 (-2.26, 1.50)	0.691

Note: Models also controlled for age, gender, country, education, employment status, and recruitment wave. The planned comparison directly contrasts the ChatGPT and AI Therapy groups. In planned comparisons, ChatGPT Group was a reference group.
[a]GAD-7 = Generalized Anxiety Disorder Scale
[b]ODSIS = Overall Depression Severity and Impairment Scale
[c]PHQ-9 = Patient Health Questionnaire
[d]WHO-5 = Well-Being Index
[e]b = Unstandardized Regression Coefficient
[f]CI = Confidence Interval

The next set of models were the planned comparisons between the AI therapy and ChatGPT (reference group) conditions, shown in the lower half of Tables 2 and 3. Although none of the estimates reached statistical significance, the AI Therapy group showed small, favorable effect size trends compared to the ChatGPT group for anxiety (d = -0.11), ODSIS depression scores (d = -0.13), PHQ depressive symptoms (d = -0.03), but not in well-being (d = -0.07). In the per-protocol results, there was a reverse trend in depression: ChatGPT group achieved a slightly higher reduction in depression scores (in both ODSIS and PHQ-9) as compared to the AI therapy group.

Table 3. Effect sizes (Cohen's d [95% CI]) for intention-to-treat intervention effects.

Effect	Anxiety (GAD-7[a]) d[e] (95% CI[f])	Depression (ODSIS[b]) d (95% CI)	Depression (PHQ-9[c]) d (95% CI)	Mental Well-being (WHO-5[d]) d (95% CI)
Interaction effects				
Control Group × Time	*Reference*			
AI Therapy Group × Time	-0.37 [-0.83, 0.08]	-0.25 [-0.66, 0.15]	-0.47 [-0.84, -0.11]	0.12 [-0.26, 0.51]
ChatGPT Group × Time	-0.27 [-0.70, 0.16]	-0.12 [-0.50, 0.26]	-0.44 [-0.82, -0.06]	0.20 [-0.17, 0.56]
Planned Comparisons				
ChatGPT Group × Time	*Reference*			
AI Therapy Group × Time	-0.11 [-0.54, 0.32]	-0.13 [-0.51, 0.25]	-0.03 [-0.46, 0.39]	-0.07 [-0.44, 0.29]

Note: Cohen's d was calculated by dividing the unstandardized regression coefficient (b) and its confidence interval by the pooled baseline standard deviation. A negative d indicates a greater reduction in symptoms for the non-reference group.

[a]GAD-7 = Generalized Anxiety Disorder Scale
[b]ODSIS = Overall Depression Severity and Impairment Scale
[c]PHQ-9 = Patient Health Questionnaire
[d]WHO-5 = Well-Being Index
[e]b = Unstandardized Regression Coefficient
[f]CI = Confidence Interval

Discussion

As digital therapy alternatives rapidly evolve, including genAI-based therapy chatbots, there is a need for careful, iterative testing of these novel therapeutic modalities. This pilot study provides early-stage tests of feasibility and proof-of-concept for a brief 3-week (9 session) AI Therapy chatbot intervention based upon solution-focused principles (i.e., ChatMind), relative to a standard untrained chatbot (GPT-4o-based models) condition and an AOC condition. As described below, findings provide foundational support for the overall promise/merits of genAI chatbot therapy while also elucidating key areas of improvement regarding feasibility.

Clinical Outcomes

Overall, both the AI therapy group and the ChatGPT group showed significant reductions in depressive symptoms and non-significant improvements in anxiety symptoms compared to the AOC. Changes in well-being scores were not significant in any group. Neither AI group significantly outperformed the other, although the AI therapy group exhibited greater descriptive changes.

Depression

In both the AI therapy group and the ChatGPT group, a significant reduction in depressive symptoms measured by the PHQ-9 was observed compared to the AOC, along with a non-significant but descriptive decline in the ODSIS score. The ODSIS, unlike the PHQ-9, which measures symptom severity, assesses the impact of symptoms on daily functioning, suggesting that AI therapy may influence symptom severity more than functional impairment. Overall, the reduction in depressive symptoms aligns with previous meta-analyses highlighting the short-term effectiveness of CAs on depression [26]. The effect sizes achieved by the AI interventions on the PHQ-9 (AI therapy: $d = -0.47$; ChatGPT: $d = -0.44$) are comparable to those reported for psychotherapies for depression, which typically yield standardized mean differences (SMDs) ranging from 0.11 to 0.61 [35]. On both measures, the AI therapy group showed greater reductions, although these differences were not statistically significant compared to the ChatGPT group. The AI therapy intervention in this study, though based on generative AI, followed structured prompts emphasizing ventilation and goal setting inspired by solution-focused brief therapy (SFBT), which has demonstrated effectiveness for depression [36]. This structured approach may have contributed to the greater descriptive improvements compared to the unstructured ChatGPT intervention.

Anxiety

Both the AI therapy group and the ChatGPT group demonstrated non-significant reductions in anxiety symptoms. A longer intervention may be required for these effects to reach statistical significance, although previous meta-analytic evidence suggests that CAs can have short-term effects on both generalized and specific anxiety [25,26]. By the nature of anxiety, the real-time availability of CAs may be one factor making them effective in the short term. Furthermore, AI's ability to detect and reframe cognitive distortions [37], which often contribute to anxiety, may help reduce its symptoms. Given AI's current restriction to verbal communication, its potential may be strongest in text-based cognitive interventions. Importantly. This remains only a potential, and it is unclear whether AI can deliver such interventions without specific prompting. Recent research even suggests that generative AI models may themselves exhibit metaphorical "state anxiety" when exposed to trauma-related narratives [38], indicating that specific prompting or other adjustments may be necessary to optimize AI responses to anxious input.

Well-being

In our study, both intervention groups showed slight, non-significant improvements in well-being, while the control group declined. The effect of CAs on well-being has been inconsistent across previous studies. Although a meta-analysis of Zhong et al. shows short-term gains in well-being after a CA intervention [26], another of He et al. does not [27]. Even though the WHO-5 questionnaire is sensitive to intervention-related changes [34], well-being is considered a relatively stable construct across the lifespan [39], which may limit the short-term impact of CAs. The therapeutic approach that is most commonly integrated in CAs, cognitive behavioral therapy (CBT) [40], appears to be more effective in reducing negative affect than in enhancing positive affect [41]. This difference could explain the more consistent effects on depression and anxiety compared to well-being. In contrast, SFBT, which underpinned the AI therapy group, has demonstrated effectiveness for both affects [42]. Because CBT is often symptom-focused, its integration into CAs may be effective for reducing targeted issues like depression or anxiety, but less so for promoting overall well-being, highlighting the potential benefit of including elements from broader therapeutic frameworks.

Comparison of Two Chatbots

A key contribution of our study is the direct comparison of a structured AI therapy chatbot with an unstructured general-purpose one. Although neither group significantly outperformed the other in any outcome, in the intent-to-treat analysis, the AI therapy group showed slightly larger reductions than the ChatGPT group, with small effect sizes favoring AI therapy for anxiety (AI therapy: $d = -0.37$ vs. ChatGPT: $d = -0.27$), ODSIS depression (AI therapy: $d = -0.25$ vs. ChatGPT: $d = -0.12$), and PHQ depression (AI therapy: $d = -0.47$ vs. ChatGPT: $d = -0.44$). However, in the per-protocol analysis, this pattern reversed for depression: the ChatGPT group achieved slightly greater reductions in PHQ-9 and ODSIS scores. This discrepancy may reflect differences in adherence or user engagement, as per-protocol analysis includes only participants who completed the intervention as intended. The results of the intent-to-treat analysis may be closer to a real-life scenario, while the per-protocol analysis is closer to the ideal situation. The differences also raises questions about which features, such as the AI model used, conversation flow, or therapeutic framing, drive effectiveness. We don't yet know what the determinants of the effectiveness of CAs are, however, recent studies suggest that prompt engineering influences their relevance, empathy, and contextual responses [31].

Factors Influencing Effectiveness

The emergence of generative AI represents a fundamental shift: unlike older retrieval-based or rule-based systems, generative models can produce original and coherent text [43]. Our findings revealed descriptive trends indicating that structured AI interactions led to greater results in depression and anxiety changes than in the ChatGPT group, raising questions about which features, such as the AI model used, conversation flow, or therapeutic framing, drive effectiveness. In line with our results, previous research suggests that prompt engineering influences the CAs relevance, empathy, and contextual responses [31].

When we were to compare our results thoroughly with previous studies, we encountered several obstacles. While our study compared two generative chatbots with a control group, most earlier studies examined rule-based or retrieval-based CAs, limiting the generalizability of their findings to current models used in our trial. Moreover, many studies fail to report these cricial technical details, and even meta-analyses do not account for these differences [25,44]. Recent evidence indicates that generative AI chatbots may achieve larger effects on mental health outcomes compared to rule-based agents [27], and perform better in giving empathetic responses and establishing the alliance between the CA and the user [26].

Systematic comparisons between different types of conversational agents are therefore essential to identify which specific features drive the CAs effectiveness.

Feasibility

The multinational sample of non-treatment-seeking adults generally reported favorable attitudes towards AI-based therapy assessed at baseline; however, the suboptimal adherence to intervention protocols may suggest waning interest once involved. Indeed, only 39% of those in the AI therapy condition fully adhered to the 9-session protocol. This pattern may reflect limited engagement among participants without acute treatment needs, and is an important consideration of our study of healthy, non-treatment-seeking samples. Notably, participants with higher baseline depression symptoms on the ODSIS scale were more likely to adhere to the full protocol, although this trend did not reach statistical significance. However, this was not the case for higher depression symptom scores measured by the PHQ-9, suggesting greater adherence in those whose symptoms of depression have already begun to interfere with their lives. The relatively rigid structure of the current protocol (three sessions per week for three weeks) may have also contributed to lower protocol adherence. In contrast, real-world applications of AI Therapy are likely to be more flexible, allowing individuals to engage when most needed and at a preferred pace. Furthermore, this pilot tested a single therapeutic approach (SFBT), which may not resonate with all users or meet a wide range of needs and preferences. The fixed protocol may have felt repetitive for some participants, further limiting sustained engagement. Looking ahead, AI Therapy tools will likely need to emulate the flexibility and adaptability of human therapists by tailoring interactions to patient preferences, symptom severity, and evolving goals to maximize engagement and therapeutic impact. Future feasibility research should focus on clinical, treatment-seeking samples, explore adaptive/flexible treatment models, and further evaluate the optimal number and pacing of AI therapy sessions.

Strengths and limitations

Our study is one of the few RCTs to investigate the effectiveness of a generative mental health chatbot. The field of artificial intelligence is rapidly advancing, and our study focuses on instruments that use the most recent AI models, underscoring its urgency and relevance. To the best of our knowledge, we are the first to compare two generative chatbots against a control group, which is the strongest aspect of our study. Furthermore, study code and data are freely available online so that findings from this study can be easily replicated.

The present study also has several limitations. The first one is a high attrition rate, which is a common challenge in short-term CA intervention studies [45]. We addressed this limitation with a second participant recruitment, but although the targeted sample size estimated by power analysis was achieved, a larger sample would still be likely needed to detect smaller effects. Second limitation is the intervention length. The second limitation relates to the nature of the generative AI on which both interventions were based, as it inherently limits control over the conversation flow and content, compared to the earlier researched rule or retrieval based chatbots. However, this also allowed for more naturalistic data and greater ecological validity. The third limitation is a lack of data on the exact length of interventions. We were not able to track the total time participants spent conversing within the AI therapy and ChatGPT groups. Nonetheless, this reflects a common limitation in studies of unsupervised digital interventions. The fourth limitation is that we did not specify in the guidelines that participants should not start any psychological treatment, only that they should not use another chatbot for therapeutic purposes.

Implications

Future research should investigate the long-term effects of AI-based psychological interventions, as most studies, including ours, assess only short-term outcomes. Extended intervention periods and follow-up assessments are needed to evaluate sustainability and rule out novelty effects. Comparative studies should also explore different AI types, delivery formats (e.g., text vs. voice), and therapeutic approaches embedded in chatbots.

Regarding implications for practice, our results underscore that not all AI-based mental health tools are equal: therapeutic outcomes may critically depend on how AI is deployed in the particular chatbot, including the use of prompt engineering, conversation design, and alignment with established therapeutic frameworks. For chatbot developers, it will be important to build on development practices as well as psychological foundations that are supported by research, and also to evaluate the effectiveness of specific chatbots through new research and identify new factors contributing to their effectiveness.

Conclusion

This study evaluated the effectiveness of a structured, generative chatbot rooted in solution-focused brief therapy, compared to a general-purpose generative AI chatbot (ChatGPT) and a no-intervention control group over three weeks. The AI therapy group showed significant reductions in anxiety and depression over time compared to the control group, while ChatGPT group showed no significant changes. These findings support the potential of structured generative AI interventions for mental health. Further comparative studies are essential to identify the specific design features and therapeutic mechanisms that contribute to the effectiveness of AI-based mental health tools.

Trial Registration:

Open Science Framework: https://osf.io/r76ef

Acknowledgement

The work was supported from ERDF/ESF project DigiWELL (No. CZ.02.01.01/00/22_008/0004583) and IGA_CMTF_2025_008. Thank you to the ChatMind app, for providing access to the app for respondents.

BK, LN, and RZ contributed to the conception of the study. BK drafted the manuscript, designed the intervention process and content, recruited participants, and distributed the assessment tools. LN set up the assessment on the online platform. JF and LN performed the statistical analysis and interpreted the results. VH was the research supervisor. All authors critically revised the manuscript and approved the final version.

Conflict of Interests

The research was conducted while BK was employed at the company that owns the app studied. However, the study was carried out by university researchers, and the company did not influence data collection, analysis, or interpretation of the results.

Abbreviations

AI: Artificial intelligence
AOC: Assessment only control
CA: Conversational agent
CBT: Cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
DMHI: Digital mental health intervention
DSM-IV: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition
GAD-7: Generalized Anxiety Disorder 7-item scale

GenAI: Generative artificial intelligence
ITT: Intention-to-treat
LLM: Large language model
MCAR: Missing completely at random
ODSIS: Overall Depression Severity and Impairment Scale
OSF: Open Science Framework
PPA: Per-protocol analysis
PHQ-9: Patient Health Questionnaire-9
RCT: Randomized controlled trial
REML: Restricted maximum likelihood
SFBT: Solution-focused brief therapy
WHO-5: World Health Organization Well-Being Index (5-item version)

Data availability statement

The data, analytical scripts and further supplementary analyses used to produce the results of this study are freely available on the Open Science Framework under the following link: https://doi.org/10.17605/OSF.IO/BPS58

References

1. GBD Results. Inst Health Metr Eval. Available from: https://vizhub.healthdata.org/gbd-results [accessed June 6, 2025]
2. Institute for Health Metrics and Evaluation. GBD Compare. Available from: http://vizhub.healthdata.org/gbd-compare [accessed Mar 18, 2025]
3. McGinty EE, Eisenberg MD. Mental Health Treatment Gap—The Implementation Problem as a Research Problem. JAMA Psychiatry 2022 Aug 1;79(8):746–747. doi: 10.1001/jamapsychiatry.2022.1468
4. Mental Health Million Project. Mental Health Has Bigger Challenges Than Stigma - Rapid Report. 2021. Available from: https://mentalstateoftheworld.report/wp-content/uploads/2021/05/Rapid-Report-2021-Help-Seeking.pdf
5. Moitra M, Santomauro D, Collins PY, Vos T, Whiteford H, Saxena S, Ferrari AJ. 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 15;19(2):e1003901. PMID:35167593
6. Peipert A, Krendl AC, Lorenzo-Luaces L. Waiting Lists for Psychotherapy and Provider Attitudes Toward Low-Intensity Treatments as Potential Interventions: Survey Study. JMIR Form Res 2022 Sept 16;6(9):e39787. doi: 10.2196/39787
7. Rastpour A, McGregor C. Predicting Patient Wait Times by Using Highly Deidentified Data in Mental Health Care: Enhanced Machine Learning Approach. JMIR Ment Health 2022 Aug 9;9(8):e38428. PMID:35943774
8. Conroy J, Lin L, Ghaness A. Why people aren't getting the care they need. Monit Psychol 2020;51(5). Available from: https://www.apa.org/monitor/2020/07/datapoint-care [accessed Dec 3, 2024]
9. Andrade LH, Alonso J, Mneimneh Z, Wells JE, Al-Hamzawi A, Borges G, Bromet E, Bruffaerts R, Girolamo G de, Graaf R de, Florescu S, Gureje O, Hinkov HR, Hu C, Huang Y, Hwang I, Jin R, Karam EG, Kovess-Masfety V, Levinson D, Matschinger H, O'Neill S, Posada-Villa J, Sagar R, Sampson NA, Sasu C, Stein D, Takeshima T, Viana MC, Xavier M, Kessler RC. Barriers to Mental Health Treatment: Results from the WHO World Mental Health (WMH) Surveys. Psychol Med 2013 Aug 9;44(6):1303. PMID:23931656
10. Coêlho BM, Santana GL, Viana MC, Wang Y-P, Andrade LH. "I don't need any treatment" – barriers to mental health treatment in the general population of a megacity. Braz J Psychiatry Associação Brasileira de Psiquiatria; 2021 Apr 30;43:590–598. doi: 10.1590/1516-4446-2020-1448
11. Torous J, Myrick KJ, 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 Mar 26;7(3):e18848. doi: 10.2196/18848
12. Aboujaoude E, Gega L, Parish MB, Hilty DM. Editorial: Digital Interventions in Mental Health: Current Status and Future Directions. Front Psychiatry Frontiers; 2020 Feb 27;11. doi: 10.3389/fpsyt.2020.00111
13. Lucas GM, Gratch J, King A, Morency L-P. It's only a computer: Virtual humans increase willingness to disclose. Comput Hum Behav 2014 Aug 1;37:94–100. doi: 10.1016/j.chb.2014.04.043
14. Omylinska-Thurston J, Aithal S, Liverpool S, Clark R, Moula Z, Wood J, Viliardos L, Rodríguez-Dorans E, Farish-Edwards F, Parsons A, Eisenstadt M, Bull M, Dubrow-Marshall L, Thurston S, Karkou V. Digital Psychotherapies for Adults Experiencing Depressive Symptoms: Systematic Review and Meta-Analysis. JMIR Ment Health 2024;11(e55500). doi: doi: 10.2196/55500

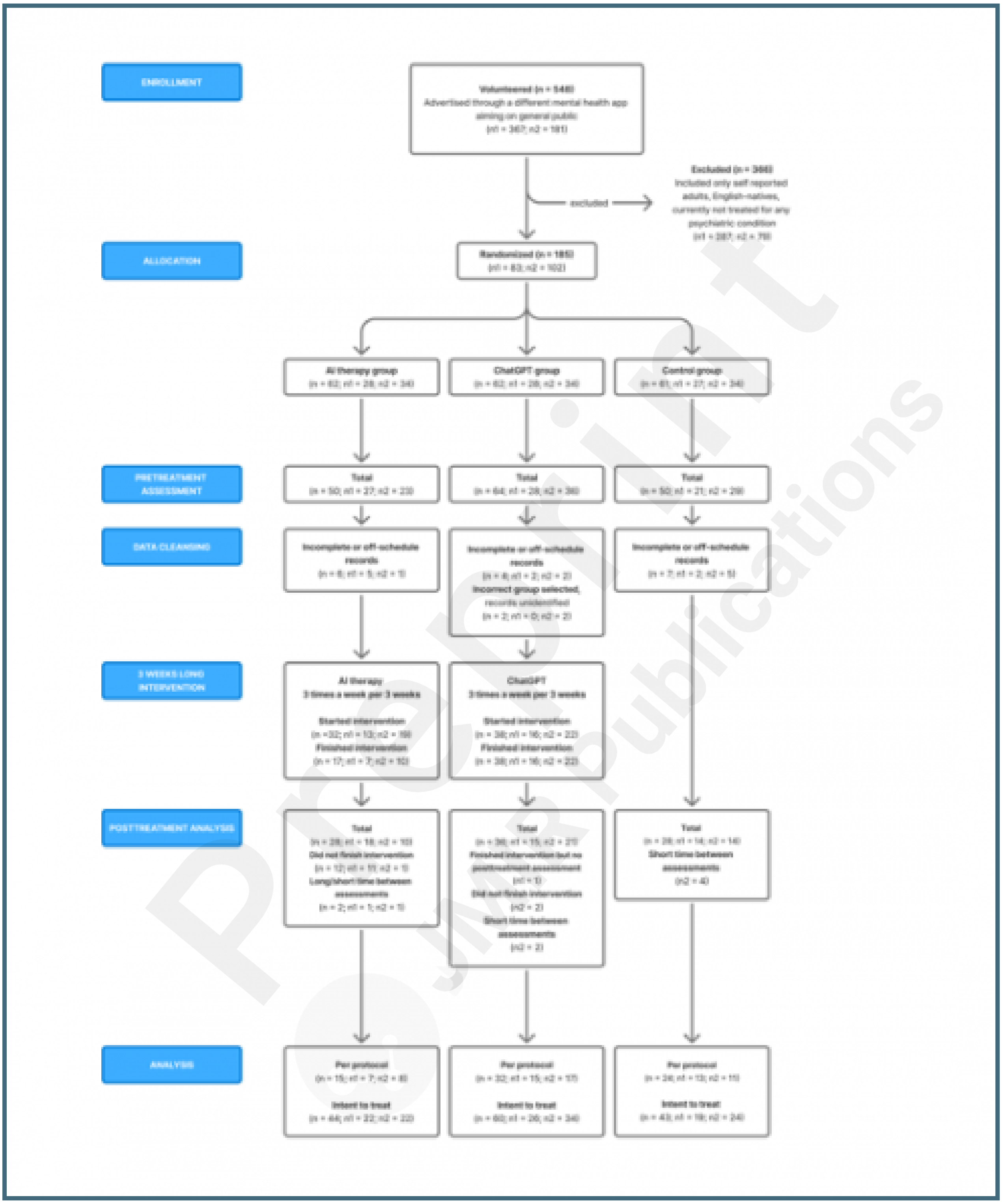
15. Pauley D, Cuijpers P, Papola D, Miguel C, Karyotaki E. Two decades of digital interventions for anxiety disorders: a systematic review and meta-analysis of treatment effectiveness. Psychol Med 2023 Jan;53(2):567–579. PMID:34047264
16. Willemsen RF, Versluis A, Aardoom JJ, Petrus AHJ, Silven AV, Chavannes NH, van Dijke A. Evaluation of completely online psychotherapy with app-support versus therapy as usual for clients with depression or anxiety disorder: A retrospective matched cohort study investigating the effectiveness, efficiency, client satisfaction, and costs. Int J Med Inf 2024 Sept 1;189:105485. doi: 10.1016/j.ijmedinf.2024.105485
17. Koulouri T, Macredie RD, Olakitan D. Chatbots to support young adults' mental health: An exploratory study of acceptability. ACM Trans Interact Intell Syst US: ACM - Association for Computing Machinery; 2022;12(2):1–39. doi: 10.1145/3485874
18. Hornstein S, Zantvoort K, Lueken U, Funk B, Hilbert K. Personalization strategies in digital mental health interventions: a systematic review and conceptual framework for depressive symptoms. Front Digit Health Frontiers; 2023 May 22;5. doi: 10.3389/fdgth.2023.1170002
19. Swift JK, Callahan JL, Cooper M, Parkin SR. The impact of accommodating client preference in psychotherapy: A meta-analysis. J Clin Psychol 2018;74(11):1924–1937. doi: 10.1002/jclp.22680
20. Kuhail MA, Alturki N, Thomas J, Alkhalifa AK, Alshardan A. Human-Human vs Human-AI Therapy: An Empirical Study. Int J Human–Computer Interact Taylor & Francis; 2024;0(0):1–12. doi: 10.1080/10447318.2024.2385001
21. Abd-alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: A scoping review. Int J Med Inf 2019 Dec 1;132:103978. doi: 10.1016/j.ijmedinf.2019.103978
22. De Freitas J, Cohen IG. The health risks of generative AI-based wellness apps. Nat Med Nature Publishing Group; 2024 May;30(5):1269–1275. doi: 10.1038/s41591-024-02943-6
23. Miner AS, Shah N, Bullock KD, Arnow BA, Bailenson J, Hancock J. Key Considerations for Incorporating Conversational AI in Psychotherapy. Front Psychiatry Frontiers; 2019 Oct 18;10. doi: 10.3389/fpsyt.2019.00746
24. Salah M, Abdelfattah F, Al Halbusi H. The good, the bad, and the GPT: Reviewing the impact of generative artificial intelligence on psychology. Curr Opin Psychol 2024 Oct 1;59:101872. doi: 10.1016/j.copsyc.2024.101872
25. Zhong W, Luo J, Zhang H. The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: A systematic review and meta-analysis. J Affect Disord 2024 July 1;356:459–469. doi: 10.1016/j.jad.2024.04.057
26. He Y, Yang L, Qian C, Li T, Su Z, Zhang Q, Hou X. Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials. J Med Internet Res 2023 Apr 28;25(1):e43862. doi: 10.2196/43862
27. Li H, Zhang R, Lee Y-C, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med 2023 Dec 19;6:236. PMID:38114588
28. Feng Y, Hang Y, Wu W, Song X, Xiao X, Dong F, Qiao Z. Effectiveness of AI-Driven Conversational Agents in Improving Mental Health Among Young People: Systematic Review and Meta-Analysis. J Med Internet Res 2025 May 14;27(1):e69639. doi: 10.2196/69639
29. Cooper ZW, Mowbray O, Ali MK, Johnson LCM. Addressing depression and comorbid health conditions through solution-focused brief therapy in an integrated care setting: a randomized clinical trial. BMC Prim Care 2024 Aug 23;25(1):313. doi: 10.1186/s12875-024-02561-8

30. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092–1097. PMID:16717171
31. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: Validity of a Brief Depression Severity Measure. J Gen Intern Med 2001 Sept;16(9):606–613. PMID:11556941
32. Bentley KH, Gallagher MW, Carl JR, Barlow DH. Development and validation of the Overall Depression Severity and Impairment Scale. Psychol Assess US: American Psychological Association; 2014;26(3):815–830. doi: 10.1037/a0036216
33. Johansen S. Wellbeing measures in primary health care/the DEPCARE project. 1998 p. 12.
34. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167–176. PMID:25831962
35. Leichsenring F, Steinert C, Rabung S, Ioannidis JPA. The efficacy of psychotherapies and pharmacotherapies for mental disorders in adults: an umbrella review and meta-analytic evaluation of recent meta-analyses. World Psychiatry 2022;21(1):133–145. doi: 10.1002/wps.20941
36. Żak AM, Pękala K. Effectiveness of solution-focused brief therapy: An umbrella review of systematic reviews and meta-analyses. Psychother Res J Soc Psychother Res 2024 Sept 26;1–13. PMID:39324877
37. Lin S, Wang Y, Dong J, Ni S. Detection and Positive Reconstruction of Cognitive Distortion sentences: Mandarin Dataset and Evaluation. arXiv; 2024. doi: 10.48550/arXiv.2405.15334
38. Ben-Zion Z, Witte K, Jagadish AK, Duek O, Harpaz-Rotem I, Khorsandian M-C, Burrer A, Seifritz E, Homan P, Schulz E, Spiller TR. Assessing and alleviating state anxiety in large language models. Npj Digit Med Nature Publishing Group; 2025 Mar 3;8(1):132. doi: 10.1038/s41746-025-01512-6
39. Vries LP de, Pelt DHM, Bartels M. The stability and change of wellbeing across the lifespan: a longitudinal twin-sibling study. Psychol Med 2024 July;54(10):2572–2584. doi: 10.1017/S0033291724000692
40. Balan R, Dobrean A, Poetar CR. Use of automated conversational agents in improving young population mental health: a scoping review. Npj Digit Med Nature Publishing Group; 2024 Mar 19;7(1):1–9. doi: 10.1038/s41746-024-01072-1
41. Whelen ML, Strunk DR. Does cognitive behavioral therapy for depression target positive affect? Examining affect and cognitive change session-to-session. J Consult Clin Psychol 2021 Sept;89(9):742–750. PMID:34591547
42. Grant AM. Making positive change: A randomized study comparing solution-focused vs. problem-focused coaching questions. J Syst Ther US: Guilford Publications; 2012;31(2):21–35. doi: 10.1521/jsyt.2012.31.2.21
43. Pandey S, Sharma S. A comparative study of retrieval-based and generative-based chatbots using Deep Learning and Machine Learning. Healthc Anal 2023 Nov;3:100198. doi: 10.1016/j.health.2023.100198
44. Lau Y, Ang WHD, Ang WW, Pang PC-I, Wong SH, Chan KS. Artificial Intelligence–Based Psychotherapeutic Intervention on Psychological Outcomes: A Meta-Analysis and Meta-Regression. Depress Anxiety 2025;2025(1):8930012. doi: 10.1155/da/8930012
45. Jabir AI, Lin X, Martinengo L, Sharp G, Theng Y-L, 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. PMID:38412023

Supplementary Files

Figures

Consort Flow Diagram.



Multimedia Appendixes

Descriptive characteristics.
URL: http://asset.jmir.pub/assets/3561218e33f2083c75c355eb4f8dd8ab.docx

Informed consent.
URL: http://asset.jmir.pub/assets/2ba4dbbc4ca1a2d28fcef497d6d049c7.docx

Instructions.
URL: http://asset.jmir.pub/assets/b4a0b3063e297d9cda8f232303138fcc.docx

ChatMind App Screenshot.
URL: http://asset.jmir.pub/assets/672fcffd40365264adcd4201214f6ecc.png

CONSORT (or other) checklists

CONSORT checklist.
URL: http://asset.jmir.pub/assets/b54330e26d680200336cf3ea095f7008.pdf