

Original Paper

Mental Health Chatbot for Young Adults With Depressive Symptoms During the COVID-19 Pandemic: Single-Blind, Three-Arm Randomized Controlled Trial

Yuhao He^{1,2}, BE, MEd; Li Yang^{1,2}, BM, MM, PhD; Xiaokun Zhu³, BEd, MEd; Bin Wu⁴, BE, ME; Shuo Zhang⁴, BE, ME; Chunlian Qian^{1,2}, BS, MEd; Tian Tian⁵, BE, ME

¹Institute of Applied Psychology, College of Education, Tianjin University, Tianjin, China

²Laboratory of Suicidology, Tianjin Municipal Education Commission, Tianjin, China

³Tianjin Vocational Institute, Tianjin, China

⁴Tianjin Quesoar Intelligent Technology Co, Ltd, Tianjin, China

⁵College of Intelligence and Computing, Tianjin University, Tianjin, China

Corresponding Author:

Li Yang, BM, MM, PhD

Institute of Applied Psychology

College of Education

Tianjin University

135 Yaguan Road, Jinnan District

Tianjin, 300354

China

Phone: 86 13752183496

Email: yangli@tju.edu.cn

Abstract

Background: Depression has a high prevalence among young adults, especially during the COVID-19 pandemic. However, mental health services remain scarce and underutilized worldwide. Mental health chatbots are a novel digital technology to provide fully automated interventions for depressive symptoms.

Objective: The purpose of this study was to test the clinical effectiveness and nonclinical performance of a cognitive behavioral therapy (CBT)-based mental health chatbot (XiaoE) for young adults with depressive symptoms during the COVID-19 pandemic.

Methods: In a single-blind, 3-arm randomized controlled trial, participants manifesting depressive symptoms recruited from a Chinese university were randomly assigned to a mental health chatbot (XiaoE; n=49), an e-book (n=49), or a general chatbot (Xiaoai; n=50) group in a ratio of 1:1:1. Participants received a 1-week intervention. The primary outcome was the reduction of depressive symptoms according to the 9-item Patient Health Questionnaire (PHQ-9) at 1 week later (T1) and 1 month later (T2). Both intention-to-treat and per-protocol analyses were conducted under analysis of covariance models adjusting for baseline data. Controlled multiple imputation and δ -based sensitivity analysis were performed for missing data. The secondary outcomes were the level of working alliance measured using the Working Alliance Questionnaire (WAQ), usability measured using the Usability Metric for User Experience-LITE (UMUX-LITE), and acceptability measured using the Acceptability Scale (AS).

Results: Participants were on average 18.78 years old, and 37.2% (55/148) were female. The mean baseline PHQ-9 score was 10.02 (SD 3.18; range 2-19). Intention-to-treat analysis revealed lower PHQ-9 scores among participants in the XiaoE group compared with participants in the e-book group and Xiaoai group at both T1 ($F_{2,136}=17.011$; $P<.001$; $d=0.51$) and T2 ($F_{2,136}=5.477$; $P=.005$; $d=0.31$). Better working alliance (WAQ; $F_{2,145}=3.407$; $P=.04$) and acceptability (AS; $F_{2,145}=4.322$; $P=.02$) were discovered with XiaoE, while no significant difference among arms was found for usability (UMUX-LITE; $F_{2,145}=0.968$; $P=.38$).

Conclusions: A CBT-based chatbot is a feasible and engaging digital therapeutic approach that allows easy accessibility and self-guided mental health assistance for young adults with depressive symptoms. A systematic evaluation of nonclinical metrics for a mental health chatbot has been established in this study. In the future, focus on both clinical outcomes and nonclinical metrics is necessary to explore the mechanism by which mental health chatbots work on patients. Further evidence is required to confirm the long-term effectiveness of the mental health chatbot via trials replicated with a longer dose, as well as exploration of its stronger efficacy in comparison with other active controls.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2100052532; <http://www.chictr.org.cn/showproj.aspx?proj=135744>

(*J Med Internet Res* 2022;24(11):e40719) doi: [10.2196/40719](https://doi.org/10.2196/40719)

KEYWORDS

chatbot; conversational agent; depression; mental health; mHealth; digital medicine; randomized controlled trial; evaluation; cognitive behavioral therapy; young adult; youth; health service; mobile health; COVID-19

Introduction

The COVID-19 pandemic has had a huge impact on people's mental health, increasing the rates of depression and anxiety by more than 25% globally in the first year, with people aged 20-24 years being more affected than older people [1]. However, there are still many limitations in traditional face-to-face psychotherapy and mental health services, including expensive treatment, geographical limitations, few experienced therapists, and delayed treatment [2], and stigma is considered as the most significant barrier to providing mental health services [3,4]. Limited accessibility and acceptability were more obvious with the rising risk of mental health problems [5,6] led by quarantine and social isolation during the COVID-19 pandemic [7], especially among adolescents [8]. Mobile health and digital medicine have rapidly become an important area of study [9] in response to the conundrum posed by the escalating demand for mental health assistance [10] and the severe shortage of traditional health care facilities [11].

Driven by digital technologies, such as computers, the internet, mobile devices, mobile software apps, and virtual reality (VR), treatment for mental health problems has undergone an unprecedented transformation [12].

A chatbot, as a novel digital technology for mental health service, is a software program that simulates conversations with users through text or voice depending on artificial intelligence (AI) [13]. The first chatbot, ELIZA, was applied in the field of psychology, and users could input text to simulate a conversation with a Rogerian psychotherapist [14]. A mental health chatbot provides more accessibility than traditional face-to-face counseling and psychotherapy [15,16], through which users can feel accompanied and understood [17,18]. In addition, chatbots were designed to focus on interactive capabilities instead of single psychological education for facilitating the process of psychotherapy [19]. Most mental health chatbots can independently provide service to users, without requiring the participation and guidance of human therapists [20]. However, studies have shown that mental health chatbots have some risks as well, such as "misunderstanding," which may lead to ineffective or even harmful interventions, lack of crisis warning mechanisms, and lack of privacy protection [21]. Chatbots in mental health are nascent [22], and although chatbots have demonstrated feasibility to provide mental health treatment, more high-quality evidence regarding the effectiveness and acceptability of mental health chatbots is needed [23], particularly during the COVID-19 pandemic [24].

According to the latest data of the World Health Organization, there were 3.22 million depressed people worldwide in 2015 [25]. In China, the figure has been reported to be 95 million

[26], and the prevalence among college students was 28.4% [27], reaching 34% during the COVID-19 pandemic [28]. However, the use of health services for depressive disorders in China has been rather limited, with the access rate of adequate treatment being less than 0.5% [29]. Cognitive behavioral therapy (CBT) has been continuously developed and is currently recognized as a widespread and effective evidence-based psychotherapy for depression [30,31], serving as one of the crucial theoretical frameworks for chatbot interventions. In recent years, a number of mental health chatbots have emerged, and their effectiveness has been tested through randomized controlled trials, providing interventions for different mental health problems, with Woebot [32], Tess [33,34], Wysa [35], Vivibot [36], and XiaoNan [37] directly targeting depression and anxiety symptoms; Shim [38], SABORI [39], and Bella [40] directly targeting stress, well-being, or quality of life; and MYLO [41-43] and Help4Mood [44] directly targeting general psychological distress such as problem solving and negative cognition.

The technology and format of chatbots for mental health problems have evolved from script bots with only text communication to embodied conversational agents [45] with image and voice, and digital humans [40] and virtual humans [46], which discern and control emotional and facial expressions during interactions with individuals in real time, have also been reported.

However, previous studies focused more on the role of a chatbot as a technical carrier in the intervention, neglecting the verification and innovation of the psychological process and content itself. As a result, there is a gap between the progress of psychology and AI in the field of digital mental health. Chatbots are intended to foster collaboration, integration, and co-development between psychological science and other fields [47]. Thus, a direct comparison between mental health chatbots and general chatbots is essential in a trial. Methodological limitations that existed in previous trials involved an insufficient sample size, a lack of follow-up assessment, failure to comprehensively investigate the long-term effectiveness of the intervention, and ignorance of the sensitivity analysis to ensure robustness of the conclusion.

As an alternative and useful precursor to clinical effectiveness, nonclinical metrics are just as important as clinical outcomes and may contribute to further exploring the mechanism by which the mental health chatbots work [48]. Fitzpatrick et al [32] also noted that therapeutic process factors of mental health chatbots may facilitate or undermine the treatment. From the technical perspective, there is currently no standard method in use to evaluate mental health chatbots. As a result, we attempted to establish a systematic evaluation of nonclinical metrics for mental health chatbots covering adherence, engagement,

working alliance, usability, acceptability, and thematic analysis of users' feedback. Working alliance (also known as "therapeutic alliance") represents the cooperative and emotionally connected relationship between the client and the therapist, and is considered a common factor in psychotherapy outcomes [49] and a metric to assess the computer-patient relationship as well [50,51]. Three recent studies [52-55] by Dosovitsky et al, Beatty et al, and Darcy et al had emphasized the viability and significance of the relationship and working alliance in digital treatment, and several randomized controlled trials of mental health chatbots had employed the Working Alliance Inventory (WAI) as a measurement method of working alliance, with all of these demonstrating good measure effects [37,56,57]. Important issues to be addressed for chatbots in the future could be extracted from the perceptions and opinions of patients [58], and thematic analysis with a topic model is a qualitative research method to accurately capture and concisely present key information in texts [59].

A randomized controlled trial including 148 Chinese college students was conducted in this study to evaluate the performance and efficacy of a mental health chatbot (XiaoE) for depression. We expected that, compared with an e-book and a general chatbot, the mental health chatbot would be more effective in reducing depressive symptoms after 1-week treatment and that this effect would persist for 1 month after the intervention (primary hypothesis). Additionally, we hypothesized that the mental health chatbot would make it easier to build relationships with users, enhance engagement, and improve user experience during the therapeutic process (secondary hypothesis).

Methods

Study Design and Participants

The study was a single-blind, 3-arm randomized controlled trial performed at a university in Tianjin, China. College students were recruited from social media outlets, online platforms, and university communities or were referred here by their counselor in the counseling center. All potential participants were screened by counseling psychologists for eligibility against the following inclusion criteria: (1) age 17-34 years; (2) average score of the depression subscale in the College Students Mental Health Screening Scale (CSMHSS) [60] within 2 to 3; and (3) ability to read Chinese. Participants were excluded if they (1) reported a score of ≥ 3 for any item in the suicide subscale in the CSMHSS; (2) reported a standard score of >3 in the suicide subscale or hallucination/delusion subscale in the CSMHSS; or (3) were taking a psychiatric medication. The CSMHSS is the mainstream tool for mental health screening of college students in China. The screening scale includes 22 dimensions that involve the main mental health problems of college students and is divided into 3 levels of screening that indicate 3 levels of mental health risk. The CSMHSS is a relevant tool for the inclusion criteria because it can not only measure the degree of depressive symptoms but also screen out individuals with high mental health risk for exclusion. Moreover, the CSMHSS is easier to implement in a university with the help of corresponding assessment platforms, given the large number of recruits. Before the enrollment, the participants were required

to carefully read and sign the written informed consent form to confirm their acceptance of the study. Participants were provided with access to artificial psychological counseling services if they had any risk of suicide, self-injury, or severe psychological distress during or after the trial, to avoid further damage. At the end of the trial, participants in control conditions were offered access to XiaoE. The trial was prospectively registered with the ChiCTR registry on October 30, 2021 (number: ChiCTR2100052532). Final data were collected on December 16, 2021. Participants received a compensation of RMB 70 (approximately US \$10) for their participation in this trial.

Randomization and Masking

Randomization with stratification by gender was performed via computer programs independently developed by the technical development team of XiaoE. Participants who were randomly assigned (1:1:1) to receive the mental health chatbot intervention, e-book intervention, or general chatbot intervention would automatically enter the corresponding intervention process when they checked into XiaoE for the first time. Treatment allocation was masked from participants, investigators, and those involved in analyzing trial data, as it was saved in an encrypted electronic file form by multiple parties (the study designer, trial implementer, data processor, and technical development representative) and unblinded after the completion of data analysis. The intervention as well as the outcome measure were completed online, and none of the investigators had access to the participants' systems during the intervention period (single blind).

Procedure

The intervention lasted for 1 week. On the day of enrollment (T0), baseline data were collected, including a pretest of the primary outcome measure (9-item Patient Health Questionnaire [PHQ-9]) and demographic information. A posttest of the primary outcome was performed 1 week later (T1), accompanied by the secondary outcomes working alliance, usability, and acceptability. A final follow-up assessment of the primary outcome was carried out 30 days after enrollment (T2).

XiaoE

XiaoE is an unguided CBT-based chatbot developed for depression, which can be used in screening, prevention, and self-assistance for depressive symptoms through a fully automatic intelligent interaction with users (text, image, and voice). The technology of XiaoE is rooted in natural language processing (NLP) and deep learning [61]. The whole chatbot dialogue system has been constructed through the open-source framework RASA [62], with content about mental health produced, discussed, and supervised by a psychologist panel led by several experienced clinical and counseling experts from schools and hospitals. XiaoE provides self-assistance service via the WeChat Official Accounts Platform. The objective of the development of XiaoE is not to replace human therapists, but to provide a convenient self-help intervention to users failing to receive immediate mental health services. It can also serve as an auxiliary tool to cooperate with traditional psychological counseling and treatment, covering functions including campus and epidemic-related counseling, adolescent mental health

screening and diagnostics, automated CBT-based chatbot interventions, intelligent multiturn conversations, artificial psychological counseling, and “tree hole” (a place to share thoughts and secrets). Participants in this condition were exposed to only the automated CBT-based chatbot intervention. Based on the principles of CBT, multiturn dialogue [63] and personalized customization were taken as the main intervention forms by referring to the content and process of several mature CBT-based chatbots [20] and internet-delivered cognitive behavioral therapy (ICBT) apps [64]. The following 7 modules were designed: “Cognition Challenge,” “Improve Self-esteem,” “Learn to Relax,” “Energy List,” “Wonderful World,” “Are You OK,” and “Escape from Loneliness,” and they correspond to the 7 concepts of psychology, cognitive distortions, self-esteem, mindfulness meditation, mental energy, natural connection, self-help, and loneliness, respectively. Participants were asked to complete a module per day in sequence during the 1-week intervention period, as well as a separate module called “Gratitude Journal” for recording positive events and mood every day.

XiaoE is equipped with complete process guidance and daily task reminders. During the implementation of the trial, the participants were only required to follow the guidance of XiaoE every day, where the staff only provided answers to technical or operational questions. In addition, the interaction data of engagement and use of XiaoE can be obtained in the background of the system. The data could not be obtained from the control groups because the interactions occurred outside the XiaoE system. As a result, the interaction frequency in the control groups was measured in the form of a self-rating questionnaire at the end of the trial.

e-Book

Participants in control group 1 were arranged to read an e-book about depression, *I Had a Black Dog* [65], which is a classic book that introduces depression knowledge to the public and guides to help depressed patients serve themselves from the first-person perspective of depressed patients and their companions. The World Health Organization adopted the animated version as its official promotional video [66] on the theme of depression. In addition, participants in the group were presented with a high-quality depression-related article daily, with the theme of each article corresponding to the daily theme of the functional modules of the intervention group.

Xiaoai

Participants in control group 2 were asked to communicate with Xiaoai at least once a day. Xiaoai is a chatbot in China designed to cater to the demands of a wider audience for small talk and not particularly for mental health services such as depression. The chat content between participants and Xiaoai was unrestricted. However, we limited the daily conversation topics (corresponding to the daily functional modules of the intervention group) and proposed specific chat tasks to the participants. For example, the topic on day 2 was self-esteem, and we endorsed that participants share their perspectives and feelings on self-esteem with Xiaoai, discuss “how self-esteem affects our emotional state and what is the relationship between

it and depression,” assess their current level of self-esteem with Xiaoai, and ask for advice on “how to improve it.”

Outcomes

Primary Outcome

The primary outcome was the score of the PHQ-9 [67], which is one of the most widely used, reliable, and validated measures of depressive symptoms. It is a 9-item self-report questionnaire that assesses the frequency and severity of depressive symptoms within the previous 2 weeks based on the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV) criteria for major depressive disorder on a 4-point scale from 0 (not at all) to 3 (nearly every day). Scores ranging from 0 to 5 indicate no symptoms of depression, and scores of 5-9, 10-14, 15-20, and 20 represent mild, moderate, moderately severe, and severe depression, respectively.

Secondary Outcomes

The secondary outcomes were the scores of the Working Alliance Questionnaire (WAQ) [68], the Usability Metric for User Experience-LITE (UMUX-LITE) [69], and the Acceptability Scale (AS). The WAQ is based on the Helping Alliance Questionnaire (HAQ-II), WAI, and California Psychotherapy Alliance Scales (CAL-PAS), with three 4-item subscales assessing the development of an affective bond in treatment and the level of agreement with treatment goals and treatment tasks. The scores of all 12 items range from 0 (rarely) to 5 (always). Usability, as “the extent to which a product can be used by specified users to achieve specific goals with effectiveness, efficiency, and satisfaction in a specified context of use” [70], was assessed by the UMUX-LITE, with 2 items to assess usefulness and ease of use, respectively, ranging from 0 (rarely) to 5 (always). Acceptability, referring to psychological acceptability for the therapeutic process and content, was assessed using a 5-point Likert scale (AS), referring to items used in previous studies on mental health chatbots [33,71] covering overall satisfaction, content satisfaction, emotional awareness, learning new knowledge, relevance to daily life, and promotion of the self-help process.

Statistical Analyses

Sample size calculation was conducted with G* Power (version 3) [72]. Latest research showed a large effect ($d=0.83$) of a chatbot intervention for depression in college students [37]. On the assumption that a replication study might be expected to achieve broadly similar results, we calculated that a sample size of 32 in each group would have 90% power to detect a net effect size of 0.83, using analysis of covariance (ANCOVA) with a 2-sided significance level of .05, while also allowing for a 20% loss to follow-up.

Difference tests were conducted with SPSS (version 26; IBM Corp). In order to determine whether any significant differences between groups existed at baseline, *F* tests with one-way analysis of variance (ANOVA) were performed on continuous baseline variables (PHQ-9 and age), and chi-square analyses were performed on categorical or nominal variables (gender, ethnicity, only child, single parent, religion, home location, and parental marriage). The same comparisons of baseline

characteristics were conducted between dropouts and participants who completed the study. Adjusted mean changes in the PHQ-9 score from baseline to T1 and T2 were analyzed as the primary efficacy endpoint using an ANCOVA model with the treatment group as the fixed effect and the corresponding baseline value as the covariate. A covariate was removed from the statistical model in case of significant interaction effects being found between this covariate and the group. A post-hoc test with Bonferroni correction was employed for multiple group comparisons. η^2 was calculated and converted to Cohen d to examine the effect size of the group difference [73]. A Cohen d of 0.2 represents a small effect; 0.5, a moderate effect; and 0.8, a large effect [74]. F tests with ANOVA were performed for the results of secondary outcomes.

The results of both the intention-to-treat (ITT) analysis [75] on the full analysis set (all enrolled participants) and the per-protocol (PP) analysis on the PP set (participants in the full analysis set without important protocol violations leading to exclusion) were reported by including all available observations in the analysis [76]. Using *mi impute* within Stata (version 15; StataCorp), we processed missing data via multiple imputation (MI) methods and performed further sensitivity analysis via δ -based methods [77].

There are 3 broad classes of missing data mechanism assumptions [78]: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MI is based on MAR, where the probability of a datum being missing does not depend on the unobserved value of the datum, but only depends on the observed values of other recorded variables. Nevertheless, missing data may not necessarily conform to MAR. Instead, they may follow MNAR, where the probability of a datum being missing does depend on the unobserved value of the datum, even given the observed data. We cannot distinguish between MNAR and both MAR and MCAR since the true values of missing data are never known, which means the results of MI may be biased. The publication of ICH E9 (R1) [79], addendum on estimands and sensitivity analysis in clinical trials, states that sensitivity analysis of missing data should be performed to ensure the robustness of the results. As a result, we performed a sensitivity analysis with δ -based methods to see if the effect remained significant when missing data followed MNAR. δ -based MI entails modifying the MAR imputation distribution using a specified numerical delta parameter to make predicted responses better or worse than predicted under MAR. For a continuous outcome, δ , the offset parameter can represent the difference in the mean response between the observed and unobserved cases [80]. Usually, the sensitivity analysis will repeat for a range of δ values corresponding to 25%, 50%, 75%, and 100% of the absolute change from baseline of outcomes in all participants.

Adherence is revealed by chi-square analysis of the attrition of participants, and engagement is revealed by the frequency and duration of the interaction with the chatbot. An interaction was considered a session if there was engagement with the chatbot lasting at least 2 user inputs within 2 minutes and a break no

longer than 1 minute. Mean interaction frequency was defined as the average number of sessions each participant had with the chatbot per day during the 1-week intervention period. Mean interaction duration was defined as the average response time of each session calculated in milliseconds between the first time the user inputs content and the last time the chatbot outputs content per day. The 1-week intervention period was divided into days 1 through 7, and each day's 24 hours were divided into 12 two-hour time periods. We recorded and calculated the mean interaction frequency and mean interaction duration for the 7 days and the 12 time periods. We recontacted all the enrolled participants after all the follow-up measurements were finished and opened access to XiaoE. They were asked 3 open-ended questions at the end of the trial: "What was your best experience using XiaoE?" "What was your worst experience using XiaoE?" and "Please make some personal comments or suggestions on XiaoE." We ran a thematic analysis on participants' feedback using Latent Dirichlet Allocation (LDA) [81], an unsupervised learning algorithm, with Pycharm (version 2020.2.2). In order to confirm the optimal number of themes for participants' feedback on each question, the perplexity under different numbers of themes should be calculated, and the topic model with the minimum perplexity should be selected. Five keywords were extracted from each theme, and each theme was named by combining keywords and original feedback text labeled as corresponding themes.

Ethics Approval

The study protocol was approved by the Medical Ethics Committee of Tianjin Anding Hospital (Tianjin Mental Health Center; number: 2021-21). All participants provided informed consent.

Results

Participant Characteristics

Figure 1 shows the participant flow (CONSORT flow diagram) [82]. A total of 379 college students were assessed for eligibility and enrolled between September 1, 2021, and November 15, 2021, of whom 143 did not meet the study criteria, 48 could not be contacted again, 19 declined to participate, 15 did not sign the written informed consent form, and 6 failed to complete the baseline measure. Ultimately, 148 participants were enrolled and randomized, of whom 49 were allocated to use the mental health chatbot (XiaoE), 49 were allocated to read the e-book, and 50 were allocated to use the general chatbot (Xiaoai). Participants were on average 18.78 years old (SD 0.89; range 17-21 years), and 37.2% (55/148) were female. The mean PHQ-9 score was 10.02 (SD 3.18; range 2-19) at baseline, just reaching the level of moderate depression. There were no significant differences in baseline characteristics among the 3 arms, as well as between dropouts and participants who completed the study (Table 1). Five participants (1 from the XiaoE group and 4 from the Xiaoai group) were identified by counseling psychologists as high risk during and after the course of the trial and underwent artificial psychological counseling.

Figure 1. Flow of participants (CONSORT). CSMHSS, College Students Mental Health Screening Scale.

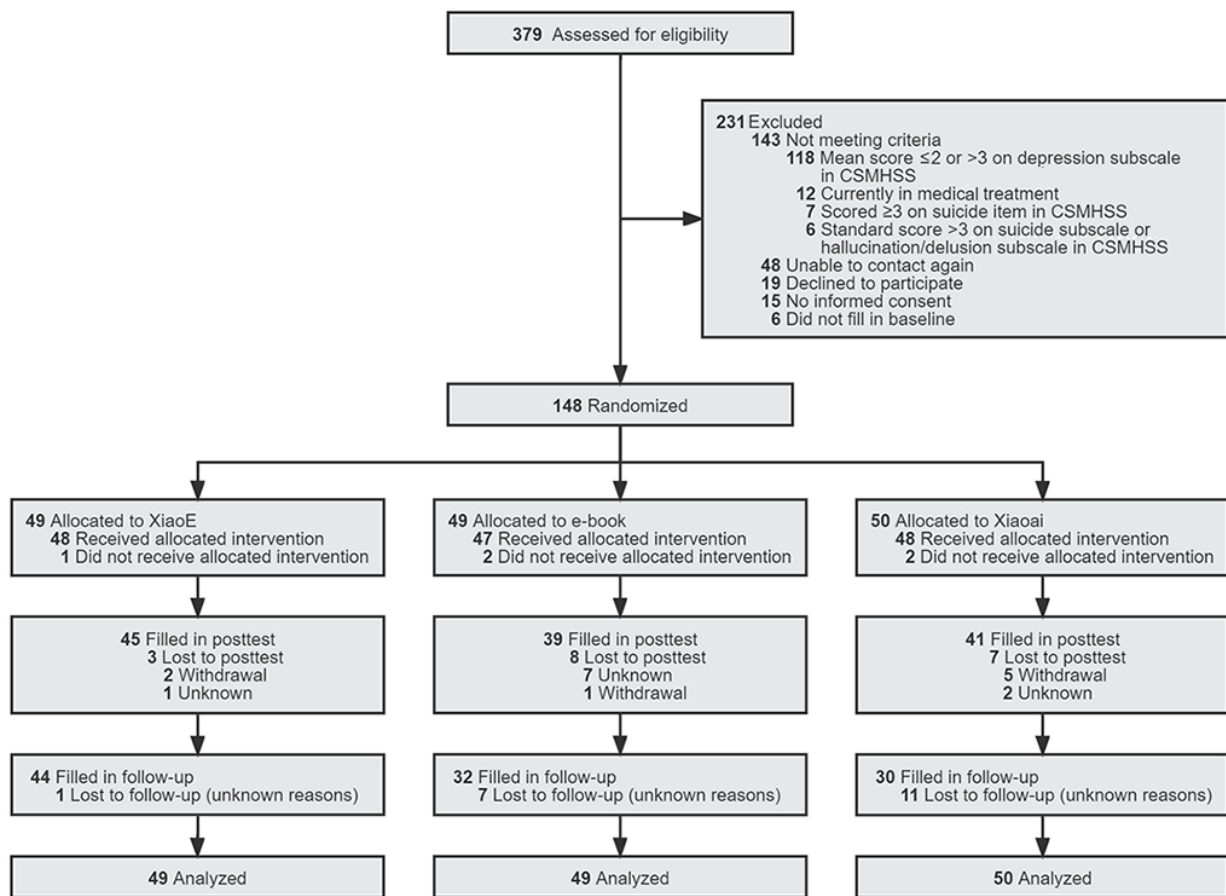


Table 1. Baseline characteristics by randomization arm.

Characteristic	XiaoE (N=49)	e-book (N=49)	Xiaoai (N=50)	Total (N=148)	F/χ^2 (df) ^b	<i>P</i> value
PHQ-9 ^a score, mean (SD)	10.10 (3.18)	9.18 (3.94)	10.76 (3.86)	10.02 (3.71)	2.294 (2,145)	.11
Age, mean (SD)	18.80 (0.89)	18.92 (0.84)	18.64 (0.90)	18.78 (0.88)	1.258 (2,145)	.29
Gender, n (%)					0.023 (2)	.99
Male	31 (63.3)	31 (63.3)	31 (62.0)	93 (62.8)		
Female	18 (36.7)	18 (36.7)	19 (38.0)	55 (37.2)		
Ethnicity, n (%)					3.239 (2)	.20
Han	44 (89.8)	44 (89.8)	49 (98.0)	137 (92.6)		
Non-Han	5 (10.2)	5 (10.2)	1 (2.0)	11 (7.4)		
Only child, n (%)					2.043 (2)	.36
Yes	16 (32.7)	13 (26.5)	10 (20.0)	39 (26.3)		
No	33 (67.3)	36 (73.5)	40 (80.0)	109 (73.7)		
Single parent, n (%)					0.450 (2)	.80
Yes	6 (12.2)	4 (8.2)	5 (10.0)	15 (10.1)		
No	43 (87.8)	45 (91.8)	45 (90.0)	133 (89.9)		
Religion, n (%)					1.912 (2)	.38
Yes	3 (6.1)	4 (8.2)	1 (2.0)	8 (5.4)		
No	46 (93.9)	45 (91.8)	49 (98.0)	140 (94.6)		
Home location, n (%)					5.057 (4)	.28
Urban	11 (22.5)	14 (28.6)	12 (24.0)	37 (25.0)		
Suburban	10 (20.4)	6 (12.2)	15 (30.0)	31 (20.9)		
Rural	28 (57.1)	29 (59.2)	23 (46.0)	80 (54.1)		
Parental marriage, n (%)					6.089 (4)	.19
Harmony	36 (73.5)	42 (85.7)	45 (90.0)	123 (83.1)		
Disharmony	7 (14.3)	5 (10.2)	2 (4.0)	14 (9.5)		
Divorced	6 (14.2)	2 (4.1)	3 (6.0)	11 (7.4)		

^aPHQ-9: 9-item Patient Health Questionnaire.

^b*F* value for PHQ-9 and age, and ² for gender, ethnicity, only child, single parent, religion, home location, and parental marriage.

Adherence and Attrition

Of the 49 participants allocated to the XiaoE group, 4 dropped out over the 1-week period and 1 dropped out over the 1-month period. Of the 49 participants allocated to the e-book group, 10 dropped out over the 1-week period and 7 dropped out over the 1-month period. Of the 50 participants allocated to the Xiaoai group, 9 dropped out over the 1-week period and 11 dropped out over the 1-month period (Figure 1). There was a lower attrition in the intervention condition compared with the control conditions (37% vs 10%; $\chi^2_1=11.904$; $P<.001$).

Effectiveness

ITT Analysis

At T1, no significant interaction effects were found between group and baseline PHQ-9 score ($P=.86$), age ($P=.91$), gender ($P=.32$), ethnicity ($P=.20$), only child ($P=.33$), single parent ($P=.99$), religion ($P=.54$), home location ($P=.62$), and parental

marriage ($P=.59$) with the ANCOVA model. Similarly, at T2, no significant interaction effects were found between group and baseline PHQ-9 score ($P=.16$), age ($P=.14$), gender ($P=.43$), ethnicity ($P=.96$), only child ($P=.27$), single parent ($P=.59$), religion ($P=.87$), home location ($P=.90$), and parental marriage ($P=.66$) with the ANCOVA model.

Depressive symptoms significantly reduced more among participants in the XiaoE group in comparison with controls, and a moderate between-group effect size was reported at T1 ($F_{2,136}=17.011$; $P<.001$; $d=0.51$), while a small effect size was reported at T2 ($F_{2,136}=5.477$; $P=.005$; $d=0.31$) (Table 2). The post-hoc test with Bonferroni correction revealed significant treatment differences with XiaoE versus e-book and Xiaoai in the reduction of depression at T1 ($P=.04$ and $P<.001$, respectively) and T2 ($P=.049$ and $P=.006$, respectively) (Figure 2).

All results were robust under sensitivity analysis, except for the comparison with e-book at T2, which changed from significant to not significant (Table 3).

Table 2. Primary outcome measures and between-group differences in the full analysis set and per-protocol set.

Analysis and timepoint	XiaoE		e-book		Xiaoai		F (df)	P value	η^2	Cohen's d
	Adjusted ^a PHQ-9 ^b , mean (SE)	n	Adjusted ^a PHQ-9 ^b , mean (SE)	n	Adjusted ^a PHQ-9 ^b , mean (SE)	n				
ITT^c analysis										
Postintervention	7.58 (0.30)	45	8.62 (0.30)	39	10.10 (0.30)	41	17.011 (2,136)	<.001	0.060	0.51
Change from baseline	-2.44 (0.30)		-1.40 (0.30)		0.08 (0.30)					
Follow-up	7.82 (0.34)	44	9.01 (0.35)	32	9.39 (0.35)	30	5.477 (2,136)	.005	0.024	0.31
Change from baseline	-2.20 (0.34)		-1.01 (0.35)		-0.63 (0.35)					
PP^d analysis										
Postintervention	7.51 (0.28)	45	9.29 (0.30)	39	10.51 (0.30)	41	26.168 (2,113)	<.001	0.088	0.62
Change from baseline	-2.84 (0.28)		-1.06 (0.30)		0.16 (0.30)					
Follow-up	7.92 (0.37)	44	9.23 (0.43)	32	10.04 (0.46)	30	6.408 (2,94)	.002	0.044	0.43
Change from baseline	-2.41 (0.37)		-1.10 (0.43)		-0.29 (0.46)					

^aAdjusted for baseline PHQ-9 score, age, gender, ethnicity, only child, single parent, religion, home location, and parental marriage.

^bPHQ-9: 9-item Patient Health Questionnaire.

^cITT: intention-to-treat.

^dPP: per-protocol.

Figure 2. Efficacy for the reduction of depression symptoms in participants. The image presents the mean change from baseline in the primary outcome measure (9-item Patient Health Questionnaire [PHQ-9]) and the between-group differences in participants with XiaoE versus those with e-book and Xiaoai at postintervention and at follow-up. Means and standard errors are displayed. (A) Intention-to-treat analysis. (B) Per-protocol analysis. * $P < .05$; ** $P < .01$; *** $P < .001$.

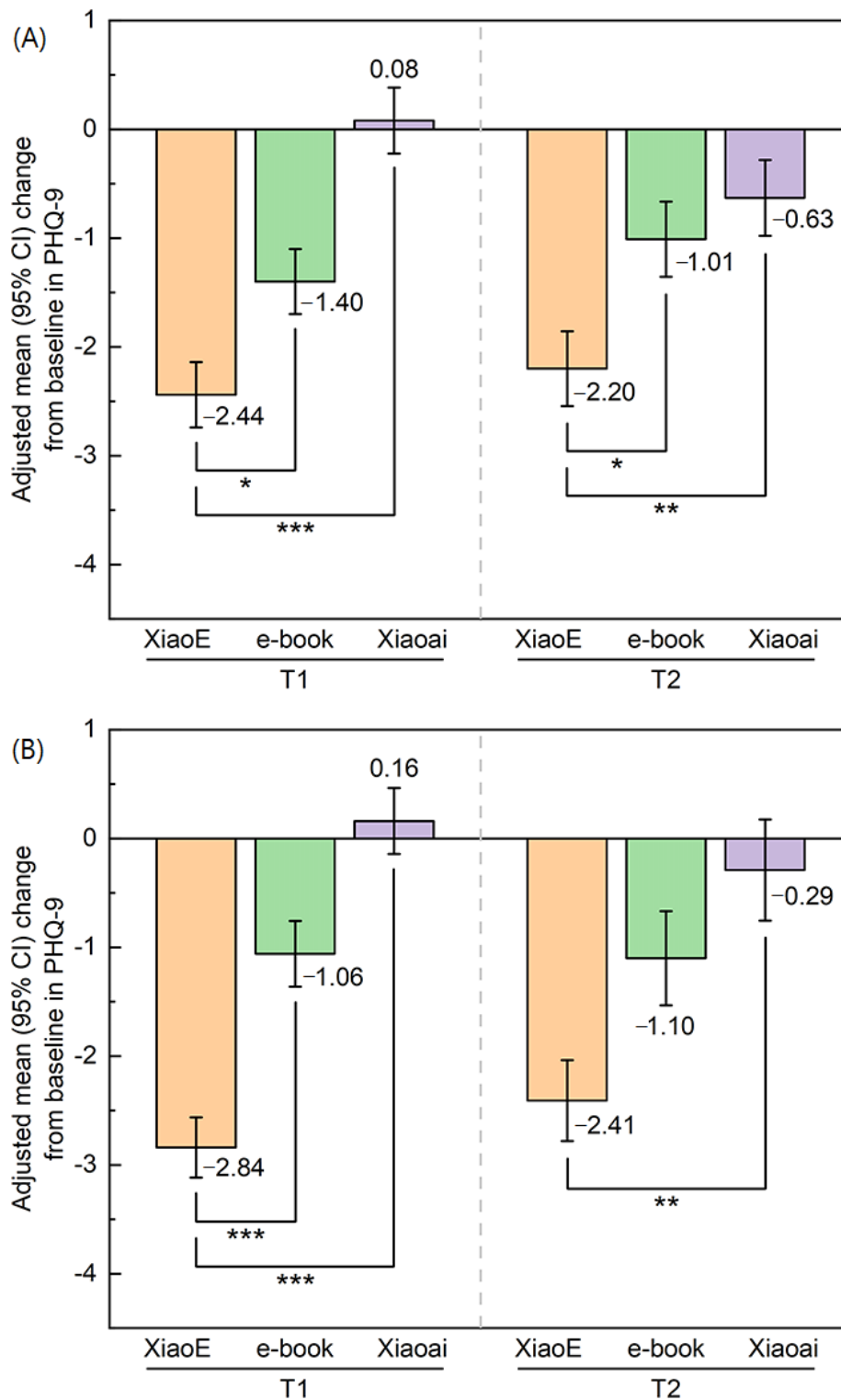


Table 3. δ -based sensitivity analysis.

Time and analysis ^a	Compared to e-book			Compared to Xiaoai		
	Group difference, value (SE)	95% CI	P value	Group difference, value (SE)	95% CI	P value
T1 (after 1 week)						
MI ^b , MAR ^c	-1.52 (0.43)	-2.38 to -0.66	.001	-2.62 (0.42)	-3.45 to -1.78	<.001
$\delta=-0.31$	-1.45 (0.43)	-2.31 to -0.59	.001	-2.52 (0.42)	-3.36 to -1.69	<.001
$\delta=-0.62$	-1.38 (0.44)	-2.25 to -0.51	.002	-2.43 (0.43)	-3.27 to -1.58	<.001
$\delta=-0.93$	-1.31 (0.44)	-2.19 to -0.43	.004	-2.33 (0.43)	-3.18 to -1.48	<.001
$\delta=-1.24$	-1.24 (0.45)	-2.13 to -0.35	.007	-2.24 (0.44)	-3.10 to -1.37	<.001
T2 (after 1 month)						
MI, MAR	-1.11 (0.54)	-2.18 to -0.03	.043	-1.65 (0.55)	-2.74 to -0.56	.003
$\delta=-0.32$	-1.03 (0.54)	-2.11 to 0.04	.06	-1.55 (0.55)	-2.64 to -0.46	.006
$\delta=-0.64$	-0.96 (0.55)	-2.05 to 0.12	.08	-1.45 (0.55)	-2.55 to -0.35	.01
$\delta=-0.96$	-0.89 (0.55)	-1.98 to 0.20	.11	-1.35 (0.56)	-2.46 to -0.24	.02
$\delta=-1.28$	-0.81 (0.56)	-1.92 to 0.29	.15	-1.25 (0.56)	-2.37 to -0.13	.03

^aThe absolute mean change from baseline to postintervention in the PHQ-9 score of all participants was -1.24, and the absolute mean change from baseline to follow-up in the PHQ-9 score of all participants was -1.28.

^bMI: multiple imputation.

^cMAR: missing at random.

PP Analysis

At T1, no significant interaction effects existed between group and baseline PHQ-9 score ($P=.59$), age ($P=.88$), gender ($P=.47$), ethnicity ($P=.44$), only child ($P=.39$), single parent ($P=.86$), religion ($P=.69$), home location ($P=.21$), and parental marriage ($P=.57$) with the ANCOVA model. Similarly, at T2, no significant interaction effects existed between group and baseline PHQ-9 score ($P=.34$), age ($P=.30$), gender ($P=.98$), ethnicity ($P=.95$), only child ($P=.11$), single parent ($P=.37$), religion ($P=.68$), home location ($P=.53$), and parental marriage ($P=.52$) with the ANCOVA model.

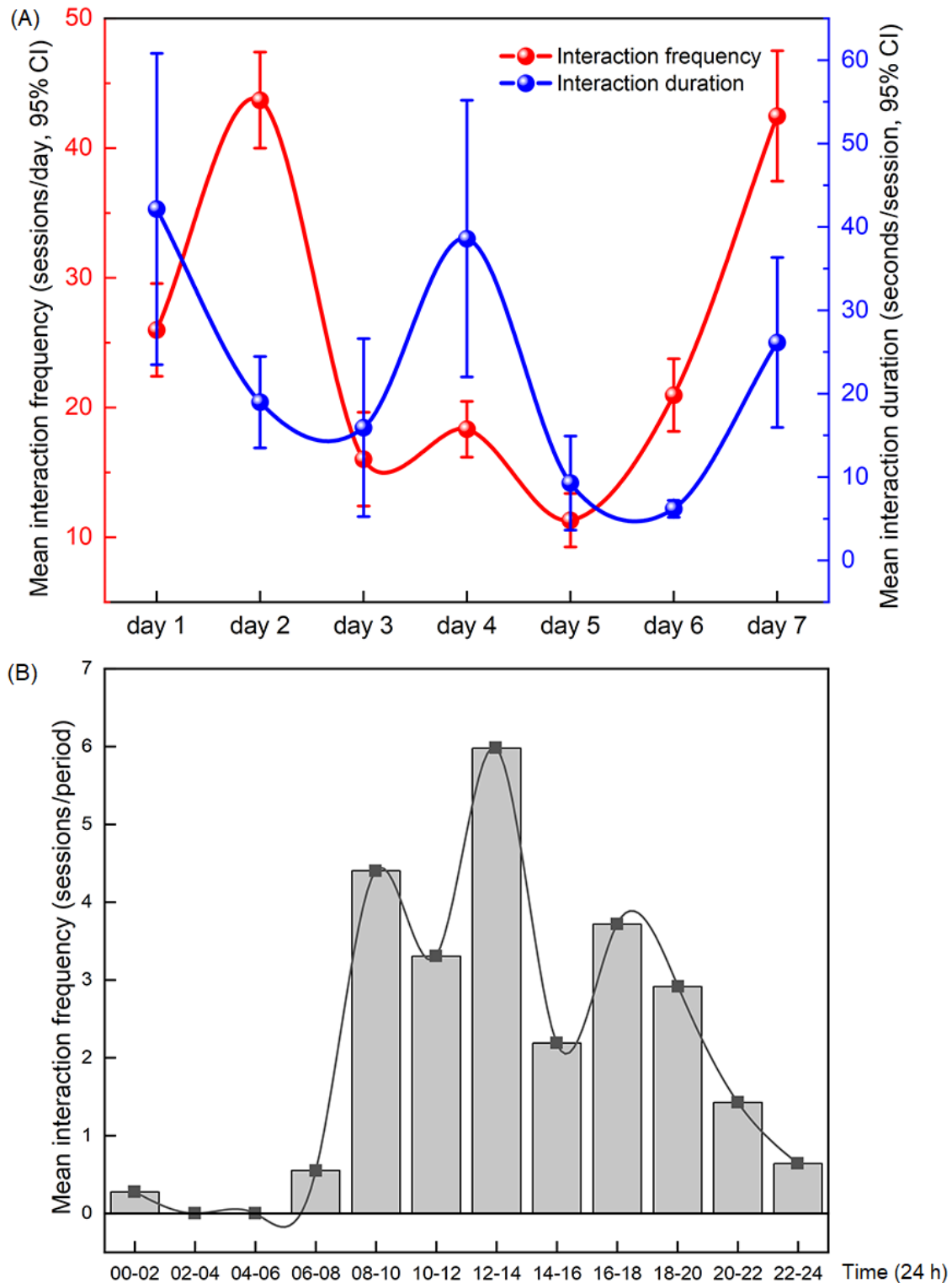
Depressive symptoms significantly reduced more among participants in the XiaoE group in comparison with controls, and a moderate between-group effect size was reported at T1 ($F_{2,113}=26.168$; $P<.001$; $d=0.62$), while a small effect size was reported at T2 ($F_{2,94}=6.408$; $P=.002$; $d=0.43$) (Table 2). The post-hoc test revealed significant treatment differences with XiaoE versus e-book and Xiaoai in the reduction of depression

at T1 ($P<.001$ and $P<.001$, respectively) and a significant difference between XiaoE and Xiaoai ($P=.003$) but no significant difference between XiaoE and e-book ($P=.08$) at T2 (Figure 2).

Use and Engagement

As shown in Figure 3, participants in the XiaoE group interacted with the chatbot for 25.54 sessions (SD 26.45; range 0-172) on average per day, and each session lasted an average of 22.46 seconds (SD 79.88; range 0-758 seconds) over the 1-week period. The daily frequency and duration of the interaction were high on day 1, day 2, and day 7, while they were relatively low on day 3, day 5, and day 6, and rebounded to some extent on day 4. The frequency of the interaction reached peaks in the 3 time periods of 8-10 AM, 12-2 PM, and 4-6 PM per day. According to the answers of participants in the e-book group, 2% (1/49) had not read it once, 51% (25/49) had read it once, and 47% (23/49) had read it twice or more. In the Xiaoai group, 29% (14/48) said they interacted with Xiaoai once a day, 27% (13/48) said twice a day, and 44% (21/48) said 3 or more times a day.

Figure 3. Use and engagement with XiaoE. The image shows the frequency and duration of interaction with the chatbot and the trend of daily interactions and interactions for 12 time periods per day in the XiaoE group during the intervention. (A) Daily engagement. The x-axis represents each day of the 1-week intervention. (B) Engagement for 12 time periods. The x-axis represents each time period in 1 day.



Working Alliance, Usability, and Acceptability

Table 4 summarizes the results of the secondary outcomes. Participants in the XiaoE condition scored higher on the total WAQ ($F_{2,145}=3.407$; $P=.04$), as well as the subscales Bond ($F_{2,145}=3.890$; $P=.02$) and Engagement ($F_{2,145}=3.925$; $P=.02$) compared with the e-book group and the Xiaoi group. No

significant difference among arms was found on the UMUX-LITE ($F_{2,145}=0.968$; $P=.38$). Better acceptability was discovered in the XiaoE group for total AS ($F_{2,145}=4.322$; $P=.02$), content satisfaction ($F_{2,145}=5.093$; $P=.007$), emotional awareness ($F_{2,145}=3.636$; $P=.03$), learning new knowledge ($F_{2,145}=4.330$; $P=.02$), and relevance to daily life ($F_{2,145}=4.834$; $P=.009$).

Table 4. Secondary outcome measures and differences between conditions.

Variable	XiaoE (n=49), mean (SD)	e-book (n=49), mean (SD)	Xiaoai (n=50), mean (SD)	F (df)	P value
WAQ^a score					
Total	53.94 (5.96)	50.35 (9.38)	50.68 (6.87)	3.407 (2,145)	.04
Goal task	17.22 (2.71)	16.43 (3.10)	16.54 (2.48)	1.188 (2,145)	.31
Bond	18.47 (1.92)	17.06 (3.26)	17.32 (2.64)	3.890 (2,145)	.02
Engagement	18.24 (2.25)	16.86 (3.54)	16.82 (2.69)	3.925 (2,145)	.02
UMUX-LITE^b score					
Total	8.61 (1.43)	8.31 (1.52)	8.24 (1.30)	0.968 (2,145)	.38
Usefulness	4.16 (0.94)	4.14 (0.76)	4.08 (0.78)	0.135 (2,145)	.87
Ease of use	4.45 (0.71)	4.16 (0.87)	4.16 (0.77)	2.192 (2,145)	.12
AS^c score					
Total	27.86 (3.25)	25.82 (5.04)	25.48 (4.53)	4.322 (2,145)	.02
Overall satisfaction	4.67 (0.75)	4.43 (0.89)	4.32 (0.89)	2.264 (2,145)	.11
Content satisfaction	4.76 (0.52)	4.45 (0.79)	4.30 (0.81)	5.093 (2,145)	.007
Emotional awareness	4.57 (0.74)	4.20 (1.00)	4.12 (0.90)	3.636 (2,145)	.03
Learning new knowledge	4.63 (0.64)	4.27 (0.95)	4.16 (0.89)	4.330 (2,145)	.02
Relevance to daily life	4.67 (0.63)	4.14 (1.10)	4.30 (0.81)	4.834 (2,145)	.009
Promotion of self-help process	4.55 (0.77)	4.33 (0.94)	4.28 (0.83)	1.429 (2,145)	.24

^aWAQ: Working Alliance Questionnaire.

^bUMUX-LITE: Usability Metric for User Experience-LITE.

^cAS: Acceptability Scale.

Thematic Analysis

According to the chart of themes-perplexity of LDA (Figure 4), the number of themes reported in the question “What was your best experience using XiaoE?” was set to 4 and the number of themes reported in the question “What was your worst experience using XiaoE?” was set to 2. Table 5 lists all the themes and keywords for participants’ feedback. The last question “comments or suggestions” was analyzed with a qualitative method because the result of LDA was not ideal.

The following 4 themes emerged in respect to the feedback to the question regarding the *best experience*: “relationship” (n=25), “emotion” (n=12), “personalization” (n=31), and “practicability” (n=80). The keywords extracted from the relationship theme were “company,” “care,” “loneliness,” “favor,” and “attending,” and the corresponding labeled example text was “XiaoE is very sweet, I like to talk to XiaoE, he will accompany and accept me, so I don’t feel lonely.” The keywords for the emotion theme were “happy,” “relax,” “stress,” “catharsis,” and “company,” and the example text was “always makes me laugh! Ha ha ha ha, the pressure suddenly disappeared, and I am so happy.” The keywords for the personalization theme were “thinking,” “learning,” “depression,”

“mood,” and “intelligence,” and the example was “The best experience is that sometimes XiaoE’s answers are indeed valuable and can really target some of my questions, which is very intelligent and promotes thinking.” The keywords for the practicability theme were “convenience,” “help,” “reality,” “method,” and “usability,” and the example was “practical, real and convenient, can help me.”

The following 2 themes emerged in respect to the feedback to the question regarding the *worst experience*: “content” (n=120) and “technology” (n=28). The keywords extracted from the content theme were “inflexible,” “response,” “tedious,” “repetitive,” and “mechanical,” and the corresponding labeled example text was “The content is too rigid. It will make people feel bored and irritable if used for a long time.” The keywords extracted from the technology theme were “glitches,” “lag,” “system,” “crash,” and “inflexible,” and the example text was “crashed when I just entered the interface, and some glitches need to be optimized.”

The participants’ feedback on the question “comments or suggestions” can be mainly extracted into the following 3 themes: hope for a more fluent process of dialogue, more emotional response and interaction, and server upgrade.

Figure 4. Chart of themes and perplexity. The image shows the perplexity under different number of themes for participant feedback of the 2 questions, "What was your best experience using XiaoE?" and "What was your worst experience using XiaoE?." (A) Themes-perplexity chart of "best experience." (B) Themes-perplexity chart of "worst experience".

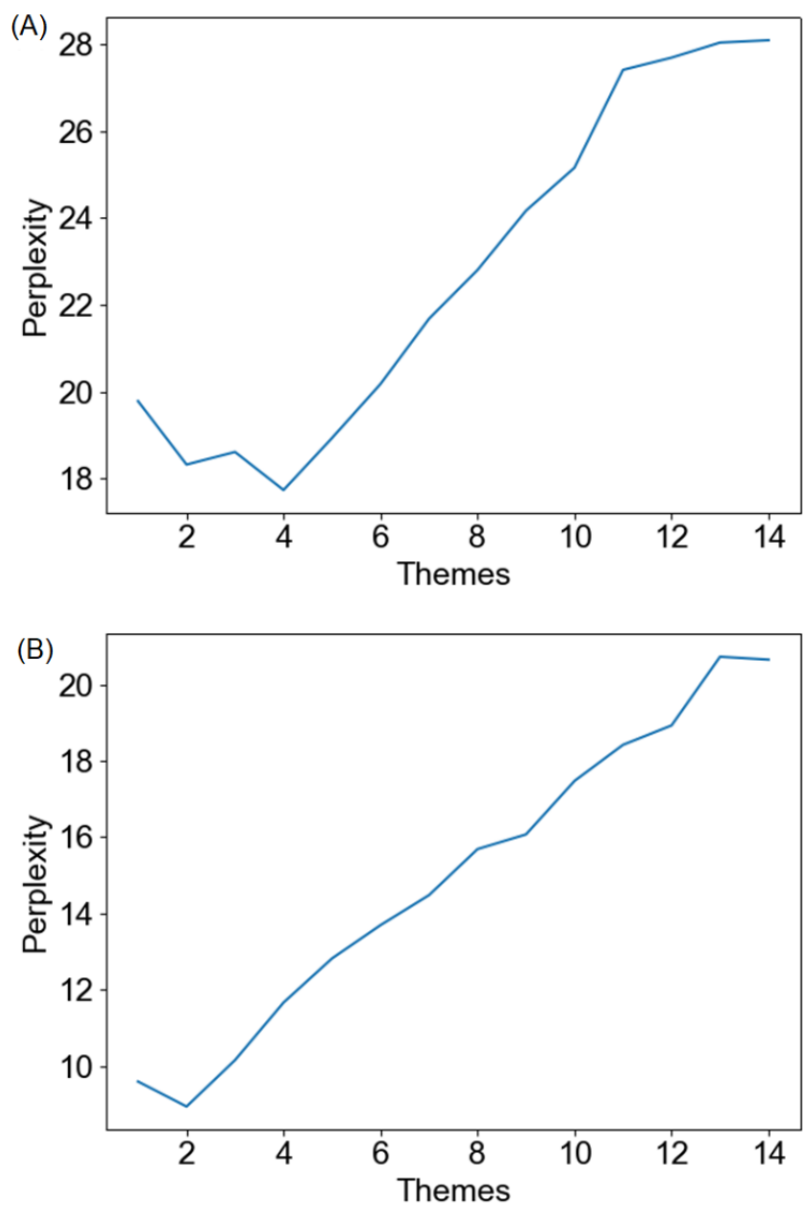


Table 5. Themes and keywords for participants' feedback.

Question and theme	Keywords
Best experience	
Relationship	Company, care, loneliness, favor, and attending
Emotion	Happy, relax, stress, catharsis, and company
Personalization	Thinking, learning, depression, mood, and intelligence
Practicability	Convenience, help, reality, method, and usability
Worst experience	
Content	Inflexible, response, tedious, repetitive, and mechanical
Technology	Glitches, lag, system, crash, and inflexible

Discussion

Principal Findings

To our knowledge, this is the first study to directly compare the clinical efficacy of a mental health chatbot with a general chatbot performing automated teletreatment for depressive symptoms. We tested both the short- and long-term effectiveness of XiaoE via a single-blind, 3-arm randomized controlled trial and established a systematic evaluation of nonclinical metrics for mental health chatbots so as to offer references for future research.

Participants in this trial were on average 18.78 years old, and they were younger than samples of typical studies with adults or college students, indicating that research on mental health chatbots is translating to samples of adolescents. In addition, more men took part in this study than in previous studies, where the majority of participants were women. Given that there are currently no well-done studies on adolescents, we hope to see more of them in the future.

In terms of attrition, participants in the XiaoE group dropped out at a lower rate than those in the e-book and Xiaoai groups. XiaoE was associated with a high level of engagement, which rose to the highest level particularly at the beginning and toward the end of the trial, demonstrating that XiaoE was attractive to participants and could quickly establish relationships when participants came into contact with this novel AI. The participants using XiaoE were most active from 12 to 2 PM every day, which may be related to the automatic task reminder set after 12 PM once a day by XiaoE. However, a large fluctuation in engagement could be observed regarding the trend of weekly activation and daily activation, which indicated that the relationships between participants and XiaoE were not steady and firm enough.

ITT analysis showed a significantly better effectiveness of XiaoE for depression in comparison with that of the 2 controls for 1 week, achieving a moderate effect size ($d=0.51$), which was between the effect sizes of 2 previous studies [32,33] (Woebot: $d=0.44$; Tess: $d=0.68$) and remained robust in sensitivity analysis. The results of the long-term reduction of depressive symptoms 1 month later were statistically significant as well, while achieving a small effect size ($d=0.31$). PP analysis also showed significant short- and long-term effectiveness (T1, $d=0.62$; T2, $d=0.43$). As in previous studies [76], the results of the ITT analysis were lower than those of the PP analysis. However, opposite results were found in separate comparisons of the XiaoE and e-book groups. The difference between XiaoE and e-book was significant in the ITT analysis (despite failing to pass the test of the sensitivity analysis), but not in the PP analysis. Protocol deviations and the interaction between compliance and the intervention, which can lead to better outcomes for compliers in the active group but just the opposite (better for noncompliers) in the control group, are commonly thought to be the causes of the bias in the PP analysis. In this study, nevertheless, the effectiveness for compliers of the e-book group may also be overestimated due to the favorable impact of compliance, which may be more significant than that in the XiaoE group. Therefore, the difference between the 2 groups

was not significant in the PP analysis. This shows that mental health chatbots should fortify the therapeutic alliance even more to increase the intervention compliance of participants.

It is necessary to note that while there was a significant improvement in symptoms via the mental health chatbot intervention, the magnitude of the improvement was small. As a result, the mental health chatbot is better suited as an auxiliary tool to work in conjunction with traditional psychological counseling and treatment or as the primary care approach for the treatment of mental illness. Although it is challenging to swiftly implement the intervention in real clinical practice, at least for the time being, the intervention is effective and convenient for individuals who desire to access self-help mental health services. This makes sense, since those represent a much larger group of people, and the spread of this unguided tool will greatly reduce the cost of human and financial resources.

XiaoE exhibited a significant high level of acceptability and work alliance with participants but a nonsignificant high level of usability. This shows that XiaoE has preliminarily reached the standard of capacity in relationship establishment, but some aspects, such as the user interface and the operating system, still need to be further simplified for users. Participants reported having received the best experience with XiaoE in the 4 themes of “relationship,” “emotion,” “personalization,” and “practicality.” The theme of “relationship” reflected the establishment, development, and function of the relationship between XiaoE and participants, as Dosovitsky et al [52] found that individuals can form a positive bond with an AI chatbot owing to its personality traits, such as being caring, open to listening, and nonjudgmental. The theme of “emotion” reflected that communication with XiaoE was helpful for emotional expression and catharsis, and made users feel accompanied and understood. The same themes were also observed in previous studies [32,33]. The theme of “personalization” reflected that XiaoE can make different suggestions for different emotional distresses put forward by the participants, which can trigger more thinking and learning of the participants. At the beginning of the content design of XiaoE, in order to avoid an overly sermonizing feeling, we added many simple and specific tips. This could be the reason why participants considered XiaoE to be practical (“practicality”). The worst experiences reported mainly focused on “content” and “technology.”

As mentioned earlier, the use of psychology in chatbots is still superficial. Despite the fact that our content was based on CBT, we discovered through our thematic analysis that participant comments barely made any mention of it. XiaoE’s conversations do not always emphasize CBT itself to participants, similar to how patients receiving therapy from a human therapist may feel like they are improving but not know what kind of therapy they are actually receiving. On the other hand, it is evident that CBT has its limitations. Even though CBT is a highly structured therapy, translating a typical CBT-based psychotherapy into a chatbot setting is difficult.

Comparison With Prior Work

We added a general chatbot (Xiaoai) intervention as a control condition to demonstrate the significance of psychological design and content for mental health chatbots. Interestingly,

participants who interacted with Xiaoi showed a small worsening of depressive symptoms after receiving the 1-week intervention. This indicates that using a general chatbot to treat mental health problems may be harmful, and a specifically designed chatbot for mental health may be required to alleviate depressive symptoms. In this study, follow-up was added to investigate the long-term effectiveness, and δ -based sensitivity analysis was performed to ensure the robustness of the conclusion. We established an innovative systematic evaluation of nonclinical metrics for mental health chatbots, and LDA was applied for the first time in the thematic analysis of users' feedback as the sample size increased.

Limitations and Future Directions

There were some limitations in this study: First, due to the particularity of the tool and the consideration of actual recruitment, it was below capacity to double-blind both the investigators and the participants. For the convenience of management, an online group was set up for the 148 enrolled participants to provide important information and technical solutions during the implementation process of the trial, which, as a potential risk, may have resulted in an attempt to reveal different contents of their own interventions, thus imposing subjective influence on the effectiveness for other participants. Special attention should be paid to this in future online research. Second, the 1-week intervention period in this study was relatively short, and the results might have shown some difference if the intervention was prolonged. It can be concluded from the trend of weekly activation that engagement with XiaoE had a wide fluctuation range and XiaoE showed strong attractiveness, but it rapidly faded in the middle of the trial. It may have resulted from repeated interactions with the inflexible and tedious content, as well as technical problems such as glitches and lag. It can be speculated that the chatbot may be more suitable for a short-term intervention rather than a long-term intervention, which needs to be explored in further studies with a longer treatment period. Third, for the control condition, the strength of evidence for the intervention itself was still limited, and the e-book intervention, as a self-help approach, only involved the concept of psychological education, and it was not equipped with a complete set of programs for psychological therapy [83] or designed for multiple or recurring sessions. Therefore, it is better to choose other active control approaches whose efficacy has been clearly proven, including traditional face-to-face therapy, online psychological counseling, ICBT, and VR. Fourth, as we could only gather self-reported involvement in the control groups as opposed to comprehensive objective data in the treatment group, it was not possible to directly compare the engagement of the XiaoE group with that of the control groups. Future research should also collect behavioral data in control groups corresponding to the data in the treatment group as the basis for comparison. Finally, this study involved students from a single university in Tianjin,

China, and it was not determined whether the conclusion can be extended to a larger group. This can be addressed by attempting to perform multicenter randomized controlled trials in the future.

In the postepidemic era, people's lifestyles have undergone profound transformations, and digital technology and internet informatization have drawn more attention than ever. It is reasonable to predict that in the future, chatbot-based digital psychotherapy will play a significant role in the field of mental health care [84]. This will provide new clinical guidelines and technical viewpoints to relevant psychologists, psychiatrists, and AI researchers and practitioners.

At present, there are many digital therapeutic approaches with excellent psychological content, with little attention to the effective factors in the psychological therapeutic process, such as emotional response, therapeutic alliance, empathy, and personalization. Despite people's doubt regarding whether machines can provide emotional experiences, they typically respond better to agents that express emotions than those that do not [85], illustrating the importance of a positive therapeutic alliance in the internet environment in the absence of therapist support [86]. Chatbots with sophisticated empathic capabilities can enrich user experience and affinity. The concept of empathic chatbots has been proposed [87], accompanied with system design and development [88], but there is no mature product present and effectiveness has not been tested yet. The utilization of user profiles or user models to support personalized and adaptive features, and assessments for personalization are still limited in mental health chatbots [89]. Thus, the technologies of chatbots, particularly NLP [61] and multiturn dialogue [60], require to be constantly upgraded, and the user interface and operating system should be modified to improve user experience. Future chatbots can be targeted at more mental health problems, such as anxiety, insomnia, well-being, stress, and addiction. Meanwhile, ethical issues with AI, such as privacy, security, information disclosure, and harm avoidance need to be carefully considered [90].

Conclusions

The mental health chatbot XiaoE can be used as a feasible, engaging, and effective digital intervention for college students with depressive symptoms. Compared with a general chatbot, XiaoE exhibited significant short-term and long-term effectiveness that remained robust after sensitivity analysis, illustrating the unique role of psychological design and process in the field of digital mental health. XiaoE showed special capacity for building relationships with users, enhancing engagement, and improving user experience during the therapeutic process. Further evidence is required to confirm the long-term effectiveness via trials replicated with a longer dose, as well as exploration of its greater efficacy in comparison with other active controls.

Acknowledgments

This study was supported by the Artificial Intelligence for Sustainable Development Goals (AI4SDGs) Research Program and the National Social Science Foundation of China (grant number: 14AZD111 and grant number: 21B[GL0031]0). We thank Xiaomi Corporation for technical support of chatbots, Tianjin Anding Hospital for supervision on the content for mental health, professor

Peng Zhang and associate professor Bo Wang for their valuable suggestions on the design and development of XiaoE, Qiang Zhang and Tong Li for their statistical expertise, and Zheng Fang and Xiaojun Hu for their efforts in the implementation of the trial.

Authors' Contributions

YH, LY, and BW contributed to the conception of the study. YH, LY, and TT designed the process and content of the XiaoE intervention. SZ contributed to the development of XiaoE. YH designed the trial and drafted the manuscript. XZ contributed to the recruitment and screening of participants. YH, XZ, CQ, and TT conducted the trial. LY, BW, and SZ commented on the design of the trial and supervised data collection and analysis. YH and CQ analyzed the data. All authors revised the content critically and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1

CONSORT-eHEALTH checklist (V 1.6.1).

[\[PDF File \(Adobe PDF File\), 1258 KB-Multimedia Appendix 1\]](#)

References

1. Mental Health and COVID-19: Early evidence of the pandemic's impact: Scientific brief, 2 March 2022. World Health Organization. URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1 [accessed 2022-05-30]
2. Christensen H. Computerised therapy for psychiatric disorders. *The Lancet* 2007 Jul;370(9582):112-113. [doi: [10.1016/S0140-6736\(07\)61065-4](https://doi.org/10.1016/S0140-6736(07)61065-4)]
3. Clement S, Schauman O, Graham T, Maggioni F, Evans-Lacko S, Bezborodovs N, et al. What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. *Psychol Med* 2015 Jan;45(1):11-27. [doi: [10.1017/S0033291714000129](https://doi.org/10.1017/S0033291714000129)] [Medline: [24569086](https://pubmed.ncbi.nlm.nih.gov/24569086/)]
4. Corrigan P. How stigma interferes with mental health care. *Am Psychol* 2004 Oct;59(7):614-625. [doi: [10.1037/0003-066X.59.7.614](https://doi.org/10.1037/0003-066X.59.7.614)] [Medline: [15491256](https://pubmed.ncbi.nlm.nih.gov/15491256/)]
5. Holmes EA, O'Connor RC, Perry VH, Tracey I, Wessely S, Arseneault L, et al. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. *Lancet Psychiatry* 2020 Jun;7(6):547-560 [FREE Full text] [doi: [10.1016/S2215-0366\(20\)30168-1](https://doi.org/10.1016/S2215-0366(20)30168-1)] [Medline: [32304649](https://pubmed.ncbi.nlm.nih.gov/32304649/)]
6. Wu T, Jia X, Shi H, Niu J, Yin X, Xie J, et al. Prevalence of mental health problems during the COVID-19 pandemic: A systematic review and meta-analysis. *J Affect Disord* 2021 Feb 15;281:91-98 [FREE Full text] [doi: [10.1016/j.jad.2020.11.117](https://doi.org/10.1016/j.jad.2020.11.117)] [Medline: [33310451](https://pubmed.ncbi.nlm.nih.gov/33310451/)]
7. Brooks SK, Webster RK, Smith LE, Woodland L, Wessely S, Greenberg N, et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *Lancet* 2020 Mar 14;395(10227):912-920 [FREE Full text] [doi: [10.1016/S0140-6736\(20\)30460-8](https://doi.org/10.1016/S0140-6736(20)30460-8)] [Medline: [32112714](https://pubmed.ncbi.nlm.nih.gov/32112714/)]
8. Loades ME, Chatburn E, Higson-Sweeney N, Reynolds S, Shafran R, Brigden A, et al. Rapid Systematic Review: The Impact of Social Isolation and Loneliness on the Mental Health of Children and Adolescents in the Context of COVID-19. *J Am Acad Child Adolesc Psychiatry* 2020 Nov;59(11):1218-1239.e3 [FREE Full text] [doi: [10.1016/j.jaac.2020.05.009](https://doi.org/10.1016/j.jaac.2020.05.009)] [Medline: [32504808](https://pubmed.ncbi.nlm.nih.gov/32504808/)]
9. Lattie EG, Adkins EC, Winkquist 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 [FREE Full text] [doi: [10.2196/12869](https://doi.org/10.2196/12869)] [Medline: [31333198](https://pubmed.ncbi.nlm.nih.gov/31333198/)]
10. Xiang Y, Yang Y, Li W, Zhang L, Zhang Q, Cheung T, et al. Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed. *The Lancet Psychiatry* 2020 Mar;7(3):228-229. [doi: [10.1016/S2215-0366\(20\)30046-8](https://doi.org/10.1016/S2215-0366(20)30046-8)]
11. Xu C, Xu Y, Xu S, Zhang Q, Liu X, Shao Y, et al. Cognitive Reappraisal and the Association Between Perceived Stress and Anxiety Symptoms in COVID-19 Isolated People. *Front Psychiatry* 2020 Sep 2;11:858 [FREE Full text] [doi: [10.3389/fpsy.2020.00858](https://doi.org/10.3389/fpsy.2020.00858)] [Medline: [32982809](https://pubmed.ncbi.nlm.nih.gov/32982809/)]
12. Fairburn CG, Patel V. The impact of digital technology on psychological treatments and their dissemination. *Behav Res Ther* 2017 Jan;88:19-25 [FREE Full text] [doi: [10.1016/j.brat.2016.08.012](https://doi.org/10.1016/j.brat.2016.08.012)] [Medline: [28110672](https://pubmed.ncbi.nlm.nih.gov/28110672/)]
13. McTear M, Callejas Z, Griol D. *The Conversational Interface: Talking to Smart Devices*. Switzerland: Springer Publishing Company; 2016.
14. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 1966 Jan;9(1):36-45. [doi: [10.1145/365153.365168](https://doi.org/10.1145/365153.365168)]

15. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Can J Psychiatry* 2019 Jul;64(7):456-464 [FREE Full text] [doi: [10.1177/0706743719828977](https://doi.org/10.1177/0706743719828977)] [Medline: [30897957](https://pubmed.ncbi.nlm.nih.gov/30897957/)]
16. Kazdin AE. Annual Research Review: Expanding mental health services through novel models of intervention delivery. *J Child Psychol Psychiatry* 2019 Apr 13;60(4):455-472. [doi: [10.1111/jcpp.12937](https://doi.org/10.1111/jcpp.12937)] [Medline: [29900543](https://pubmed.ncbi.nlm.nih.gov/29900543/)]
17. Ho A, Hancock J, Miner A. Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. *J Commun* 2018 Aug;68(4):712-733 [FREE Full text] [doi: [10.1093/joc/jqy026](https://doi.org/10.1093/joc/jqy026)] [Medline: [30100620](https://pubmed.ncbi.nlm.nih.gov/30100620/)]
18. Bickmore TW, Picard RW. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact* 2005 Jun 01;12(2):293-327. [doi: [10.1145/1067860.1067867](https://doi.org/10.1145/1067860.1067867)]
19. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc* 2018 Sep 01;25(9):1248-1258 [FREE Full text] [doi: [10.1093/jamia/ocy072](https://doi.org/10.1093/jamia/ocy072)] [Medline: [30010941](https://pubmed.ncbi.nlm.nih.gov/30010941/)]
20. Gaffney H, Mansell W, Tai S. Conversational Agents in the Treatment of Mental Health Problems: Mixed-Method Systematic Review. *JMIR Ment Health* 2019 Oct 18;6(10):e14166 [FREE Full text] [doi: [10.2196/14166](https://doi.org/10.2196/14166)] [Medline: [31628789](https://pubmed.ncbi.nlm.nih.gov/31628789/)]
21. Xu B, Zhuang Z. Survey on psychotherapy chatbots. *Concurrency and Computation* 2020 Dec 28;34(7):e6170 [FREE Full text] [doi: [10.1002/cpe.6170](https://doi.org/10.1002/cpe.6170)]
22. 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 Inform* 2019 Dec;132:103978. [doi: [10.1016/j.ijmedinf.2019.103978](https://doi.org/10.1016/j.ijmedinf.2019.103978)] [Medline: [31622850](https://pubmed.ncbi.nlm.nih.gov/31622850/)]
23. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. *J Med Internet Res* 2020 Jul 13;22(7):e16021 [FREE Full text] [doi: [10.2196/16021](https://doi.org/10.2196/16021)] [Medline: [32673216](https://pubmed.ncbi.nlm.nih.gov/32673216/)]
24. Miner AS, Laranjo L, Kocaballi AB. Chatbots in the fight against the COVID-19 pandemic. *NPJ Digit Med* 2020 May 4;3(1):65 [FREE Full text] [doi: [10.1038/s41746-020-0280-0](https://doi.org/10.1038/s41746-020-0280-0)] [Medline: [32377576](https://pubmed.ncbi.nlm.nih.gov/32377576/)]
25. Depression and Other Common Mental Disorders. World Health Organization. 2017. URL: <https://www.who.int/publications/i/item/depression-global-health-estimates> [accessed 2022-03-15]
26. Huang Y, Wang Y, Wang H, Liu Z, Yu X, Yan J, et al. Prevalence of mental disorders in China: a cross-sectional epidemiological study. *The Lancet Psychiatry* 2019 Mar;6(3):211-224 [FREE Full text] [doi: [10.1016/S2215-0366\(18\)30511-X](https://doi.org/10.1016/S2215-0366(18)30511-X)]
27. Gao L, Xie Y, Jia C, Wang W. Prevalence of depression among Chinese university students: a systematic review and meta-analysis. *Sci Rep* 2020 Sep 28;10(1):15897 [FREE Full text] [doi: [10.1038/s41598-020-72998-1](https://doi.org/10.1038/s41598-020-72998-1)] [Medline: [32985593](https://pubmed.ncbi.nlm.nih.gov/32985593/)]
28. Deng J, Zhou F, Hou W, Silver Z, Wong CY, Chang O, et al. The prevalence of depressive symptoms, anxiety symptoms and sleep disturbance in higher education students during the COVID-19 pandemic: A systematic review and meta-analysis. *Psychiatry Res* 2021 Jul;301:113863 [FREE Full text] [doi: [10.1016/j.psychres.2021.113863](https://doi.org/10.1016/j.psychres.2021.113863)] [Medline: [33984824](https://pubmed.ncbi.nlm.nih.gov/33984824/)]
29. Lu J, Xu X, Huang Y, Li T, Ma C, Xu G, et al. Prevalence of depressive disorders and treatment in China: a cross-sectional epidemiological study. *The Lancet Psychiatry* 2021 Nov;8(11):981-990. [doi: [10.1016/s2215-0366\(21\)00251-0](https://doi.org/10.1016/s2215-0366(21)00251-0)]
30. 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](https://doi.org/10.1016/j.cpr.2005.07.003)] [Medline: [16199119](https://pubmed.ncbi.nlm.nih.gov/16199119/)]
31. Malhi GS, Mann JJ. Depression. *The Lancet* 2018 Nov;392(10161):2299-2312. [doi: [10.1016/s0140-6736\(18\)31948-2](https://doi.org/10.1016/s0140-6736(18)31948-2)]
32. Fitzpatrick KK, Darcy A, Vierhile M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment Health* 2017 Jun 06;4(2):e19 [FREE Full text] [doi: [10.2196/mental.7785](https://doi.org/10.2196/mental.7785)] [Medline: [28588005](https://pubmed.ncbi.nlm.nih.gov/28588005/)]
33. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. *JMIR Ment Health* 2018 Dec 13;5(4):e64 [FREE Full text] [doi: [10.2196/mental.9782](https://doi.org/10.2196/mental.9782)] [Medline: [30545815](https://pubmed.ncbi.nlm.nih.gov/30545815/)]
34. Klos MC, Escoredo M, Joerin A, Lemos VN, Rauws M, Bunge EL. Artificial Intelligence-Based Chatbot for Anxiety and Depression in University Students: Pilot Randomized Controlled Trial. *JMIR Form Res* 2021 Aug 12;5(8):e20678 [FREE Full text] [doi: [10.2196/20678](https://doi.org/10.2196/20678)] [Medline: [34092548](https://pubmed.ncbi.nlm.nih.gov/34092548/)]
35. Inkster B, Sarda S, Subramanian V. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR Mhealth Uhealth* 2018 Nov 23;6(11):e12106 [FREE Full text] [doi: [10.2196/12106](https://doi.org/10.2196/12106)] [Medline: [30470676](https://pubmed.ncbi.nlm.nih.gov/30470676/)]
36. Greer S, Ramo D, Chang Y, Fu M, Moskowitz J, Haritatos J. Use of the Chatbot "Vivibot" to Deliver Positive Psychology Skills and Promote Well-Being Among Young People After Cancer Treatment: Randomized Controlled Feasibility Trial. *JMIR Mhealth Uhealth* 2019 Oct 31;7(10):e15018 [FREE Full text] [doi: [10.2196/15018](https://doi.org/10.2196/15018)] [Medline: [31674920](https://pubmed.ncbi.nlm.nih.gov/31674920/)]
37. Liu H, Peng H, Song X, Xu C, Zhang M. Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness. *Internet Interv* 2022 Mar;27:100495 [FREE Full text] [doi: [10.1016/j.invent.2022.100495](https://doi.org/10.1016/j.invent.2022.100495)] [Medline: [35059305](https://pubmed.ncbi.nlm.nih.gov/35059305/)]
38. Ly KH, Ly A, Andersson G. A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. *Internet Interv* 2017 Dec;10:39-46 [FREE Full text] [doi: [10.1016/j.invent.2017.10.002](https://doi.org/10.1016/j.invent.2017.10.002)] [Medline: [30135751](https://pubmed.ncbi.nlm.nih.gov/30135751/)]

39. Suganuma S, Sakamoto D, Shimoyama H. An Embodied Conversational Agent for Unguided Internet-Based Cognitive Behavior Therapy in Preventative Mental Health: Feasibility and Acceptability Pilot Trial. *JMIR Ment Health* 2018 Jul 31;5(3):e10454 [FREE Full text] [doi: [10.2196/10454](https://doi.org/10.2196/10454)] [Medline: [30064969](https://pubmed.ncbi.nlm.nih.gov/30064969/)]
40. Loveys K, Sagar M, Pickering I, Broadbent E. A Digital Human for Delivering a Remote Loneliness and Stress Intervention to At-Risk Younger and Older Adults During the COVID-19 Pandemic: Randomized Pilot Trial. *JMIR Ment Health* 2021 Nov 08;8(11):e31586 [FREE Full text] [doi: [10.2196/31586](https://doi.org/10.2196/31586)] [Medline: [34596572](https://pubmed.ncbi.nlm.nih.gov/34596572/)]
41. Bennion MR, Hardy GE, Moore RK, Kellett S, Millings A. Usability, Acceptability, and Effectiveness of Web-Based Conversational Agents to Facilitate Problem Solving in Older Adults: Controlled Study. *J Med Internet Res* 2020 May 27;22(5):e16794 [FREE Full text] [doi: [10.2196/16794](https://doi.org/10.2196/16794)] [Medline: [32384055](https://pubmed.ncbi.nlm.nih.gov/32384055/)]
42. Bird T, Mansell W, Wright J, Gaffney H, Tai S. Manage Your Life Online: A Web-Based Randomized Controlled Trial Evaluating the Effectiveness of a Problem-Solving Intervention in a Student Sample. *Behav Cogn Psychother* 2018 Sep;46(5):570-582. [doi: [10.1017/S1352465817000820](https://doi.org/10.1017/S1352465817000820)] [Medline: [29366432](https://pubmed.ncbi.nlm.nih.gov/29366432/)]
43. Gaffney H, Mansell W, Edwards R, Wright J. Manage Your Life Online (MYLO): a pilot trial of a conversational computer-based intervention for problem solving in a student sample. *Behav Cogn Psychother* 2014 Nov;42(6):731-746. [doi: [10.1017/S135246581300060X](https://doi.org/10.1017/S135246581300060X)] [Medline: [23899405](https://pubmed.ncbi.nlm.nih.gov/23899405/)]
44. Burton C, Szentagotai Tatar A, McKinstry B, Matheson C, Matu S, Moldovan R, Help4Mood Consortium. 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](https://doi.org/10.1177/1357633X15609793)] [Medline: [26453910](https://pubmed.ncbi.nlm.nih.gov/26453910/)]
45. Gardiner P, Bickmore T, Yinusa-Nyahkoon L, Reichert M, Julce C, Sidduri N, et al. Using Health Information Technology to Engage African American Women on Nutrition and Supplement Use During the Preconception Period. *Front Endocrinol (Lausanne)* 2020;11:571705 [FREE Full text] [doi: [10.3389/fendo.2020.571705](https://doi.org/10.3389/fendo.2020.571705)] [Medline: [33584534](https://pubmed.ncbi.nlm.nih.gov/33584534/)]
46. Ma T, Sharifi H, Chattopadhyay D. Virtual Humans in Health-Related Interventions: A Meta-Analysis. In: CHI EA '19: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: 2019 CHI Conference on Human Factors in Computing Systems; May 4-9, 2019; Glasgow, Scotland, UK p. 1-6. [doi: [10.1145/3290607.3312853](https://doi.org/10.1145/3290607.3312853)]
47. Lim SM, Shiao CWC, Cheng LJ, Lau Y. Chatbot-Delivered Psychotherapy for Adults With Depressive and Anxiety Symptoms: A Systematic Review and Meta-Regression. *Behav Ther* 2022 Mar;53(2):334-347. [doi: [10.1016/j.beth.2021.09.007](https://doi.org/10.1016/j.beth.2021.09.007)] [Medline: [35227408](https://pubmed.ncbi.nlm.nih.gov/35227408/)]
48. Abd-Alrazaq A, Safi Z, Alajlani M, Warren J, Househ M, Denecke K. Technical Metrics Used to Evaluate Health Care Chatbots: Scoping Review. *J Med Internet Res* 2020 Jun 05;22(6):e18301 [FREE Full text] [doi: [10.2196/18301](https://doi.org/10.2196/18301)] [Medline: [32442157](https://pubmed.ncbi.nlm.nih.gov/32442157/)]
49. Cuijpers P, Reijnders M, Huibers MJ. The Role of Common Factors in Psychotherapy Outcomes. *Annu Rev Clin Psychol* 2019 May 07;15(1):207-231. [doi: [10.1146/annurev-clinpsy-050718-095424](https://doi.org/10.1146/annurev-clinpsy-050718-095424)] [Medline: [30550721](https://pubmed.ncbi.nlm.nih.gov/30550721/)]
50. Abdulrahman A, Richards D. Modelling Therapeutic Alliance using a User-aware Explainable Embodied Conversational Agent to Promote Treatment Adherence. In: IVA '19: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. 2019 Presented at: 19th ACM International Conference on Intelligent Virtual Agents; July 2-5, 2019; Paris, France p. 248-251. [doi: [10.1145/3308532.3329413](https://doi.org/10.1145/3308532.3329413)]
51. Bickmore T, Gruber A, Picard R. Establishing the computer-patient working alliance in automated health behavior change interventions. *Patient Educ Couns* 2005 Oct;59(1):21-30. [doi: [10.1016/j.pec.2004.09.008](https://doi.org/10.1016/j.pec.2004.09.008)] [Medline: [16198215](https://pubmed.ncbi.nlm.nih.gov/16198215/)]
52. Dosovitsky G, Bunge EL. Bonding With Bot: User Feedback on a Chatbot for Social Isolation. *Front Digit Health* 2021 Oct 6;3:735053 [FREE Full text] [doi: [10.3389/fdgh.2021.735053](https://doi.org/10.3389/fdgh.2021.735053)] [Medline: [34713203](https://pubmed.ncbi.nlm.nih.gov/34713203/)]
53. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the Therapeutic Alliance With a Free-Text CBT Conversational Agent (Wysa): A Mixed-Methods Study. *Front Digit Health* 2022 Apr 11;4:847991 [FREE Full text] [doi: [10.3389/fdgh.2022.847991](https://doi.org/10.3389/fdgh.2022.847991)] [Medline: [35480848](https://pubmed.ncbi.nlm.nih.gov/35480848/)]
54. Darcy A, Daniels J, Salinger D, Wicks P, Robinson A. Evidence of Human-Level Bonds Established With a Digital Conversational Agent: Cross-sectional, Retrospective Observational Study. *JMIR Form Res* 2021 May 11;5(5):e27868 [FREE Full text] [doi: [10.2196/27868](https://doi.org/10.2196/27868)] [Medline: [33973854](https://pubmed.ncbi.nlm.nih.gov/33973854/)]
55. Nißen M, Rüeßger D, Stieger M, Flückiger C, Allemann M, V Wangenheim F, et al. The Effects of Health Care Chatbot Personas With Different Social Roles on the Client-Chatbot Bond and Usage Intentions: Development of a Design Codebook and Web-Based Study. *J Med Internet Res* 2022 Apr 27;24(4):e32630 [FREE Full text] [doi: [10.2196/32630](https://doi.org/10.2196/32630)] [Medline: [35475761](https://pubmed.ncbi.nlm.nih.gov/35475761/)]
56. Prochaska JJ, Vogel EA, Chieng A, Baiocchi M, Maglalang DD, Pajarito S, et al. A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. *Drug Alcohol Depend* 2021 Oct 01;227:108986 [FREE Full text] [doi: [10.1016/j.drugalcdep.2021.108986](https://doi.org/10.1016/j.drugalcdep.2021.108986)] [Medline: [34507061](https://pubmed.ncbi.nlm.nih.gov/34507061/)]
57. Hauser-Ulrich S, Künzli H, Meier-Peterhans D, Kowatsch T. A Smartphone-Based Health Care Chatbot to Promote Self-Management of Chronic Pain (SELMA): Pilot Randomized Controlled Trial. *JMIR Mhealth Uhealth* 2020 Apr 03;8(4):e15806 [FREE Full text] [doi: [10.2196/15806](https://doi.org/10.2196/15806)] [Medline: [32242820](https://pubmed.ncbi.nlm.nih.gov/32242820/)]

58. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and Opinions of Patients About Mental Health Chatbots: Scoping Review. *J Med Internet Res* 2021 Jan 13;23(1):e17828 [FREE Full text] [doi: [10.2196/17828](https://doi.org/10.2196/17828)] [Medline: [33439133](https://pubmed.ncbi.nlm.nih.gov/33439133/)]
59. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimed Tools Appl* 2018 Nov 28;78(11):15169-15211. [doi: [10.1007/s11042-018-6894-4](https://doi.org/10.1007/s11042-018-6894-4)]
60. Yi X, Ding X, Xue J, Wu X. Optimization of the explanation of the results of chinese college students' mental health screening scale. *Chinese Journal of Clinical Psychology* 2021;29(5):962-966 [FREE Full text]
61. Otter DW, Medina JR, Kalita JK. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Trans. Neural Netw. Learning Syst* 2021 Feb;32(2):604-624. [doi: [10.1109/tnnls.2020.2979670](https://doi.org/10.1109/tnnls.2020.2979670)]
62. Bocklisch T, Faulkner J, Pawlowski N. Rasa: Open Source Language Understanding and Dialogue Management. arXiv. 2017. URL: <https://arxiv.org/abs/1712.05181> [accessed 2022-11-11]
63. Kao C, Chen C, Tsai Y. Model of Multi-turn Dialogue in Emotional Chatbot. 2019 Presented at: 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI); November 21-23, 2019; Kaohsiung, Taiwan. [doi: [10.1109/taai48200.2019.8959855](https://doi.org/10.1109/taai48200.2019.8959855)]
64. Karyotaki E, Efthimiou O, Miguel C, Bermphohl FMG, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-Based Cognitive Behavioral Therapy for Depression: A Systematic Review and Individual Patient Data Network Meta-analysis. *JAMA Psychiatry* 2021 Apr 01;78(4):361-371 [FREE Full text] [doi: [10.1001/jamapsychiatry.2020.4364](https://doi.org/10.1001/jamapsychiatry.2020.4364)] [Medline: [33471111](https://pubmed.ncbi.nlm.nih.gov/33471111/)]
65. Johnstone M. *I Had a Black Dog*. Australia: Constable & Robinson; 2007.
66. Depression. World Health Organization. URL: <https://www.who.int/health-topics/depression> [accessed 2021-07-30]
67. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001 Sep;16(9):606-613 [FREE Full text] [doi: [10.1046/j.1525-1497.2001.016009606.x](https://doi.org/10.1046/j.1525-1497.2001.016009606.x)] [Medline: [11556941](https://pubmed.ncbi.nlm.nih.gov/11556941/)]
68. Xu Z, Guangrong J. Development of the Working Alliance Questionnaire. *Chinese Journal of Clinical Psychology* 2011(4):449-453 [FREE Full text]
69. Borsci S, Buckle P, Walne S. Is the LITE version of the usability metric for user experience (UMUX-LITE) a reliable tool to support rapid assessment of new healthcare technology? *Appl Ergon* 2020 Apr;84:103007. [doi: [10.1016/j.apergo.2019.103007](https://doi.org/10.1016/j.apergo.2019.103007)] [Medline: [31785449](https://pubmed.ncbi.nlm.nih.gov/31785449/)]
70. ISO 9241-11:2018(en) Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts. International Organization for Standardization. URL: <https://www.iso.org/obp/ui/fr/#iso:std:iso:9241:-11:ed-2:v1:en> [accessed 2022-04-10]
71. Jang S, Kim J, Kim S, Hong J, Kim S, Kim E. Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: A development and feasibility/usability study. *Int J Med Inform* 2021 Jun;150:104440. [doi: [10.1016/j.ijmedinf.2021.104440](https://doi.org/10.1016/j.ijmedinf.2021.104440)] [Medline: [33799055](https://pubmed.ncbi.nlm.nih.gov/33799055/)]
72. Faul F, Erdfelder E, Lang A, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 2007 May;39(2):175-191. [doi: [10.3758/BF03193146](https://doi.org/10.3758/BF03193146)]
73. Computation of Effect Sizes. *Psychometrica*. URL: https://www.psychometrica.de/effect_size.html [accessed 2021-12-02]
74. Cohen J. A power primer. *Psychological Bulletin* 1992;112(1):155-159. [doi: [10.1037/0033-2909.112.1.155](https://doi.org/10.1037/0033-2909.112.1.155)]
75. White IR, Horton NJ, Carpenter J, Pocock SJ. Strategy for intention to treat analysis in randomised trials with missing outcome data. *BMJ* 2011 Feb 07;342(feb07 1):d40-d40 [FREE Full text] [doi: [10.1136/bmj.d40](https://doi.org/10.1136/bmj.d40)] [Medline: [21300711](https://pubmed.ncbi.nlm.nih.gov/21300711/)]
76. Porta N, Bonet C, Cobo E. Discordance between reported intention-to-treat and per protocol analyses. *J Clin Epidemiol* 2007 Jul;60(7):663-669. [doi: [10.1016/j.jclinepi.2006.09.013](https://doi.org/10.1016/j.jclinepi.2006.09.013)] [Medline: [17573981](https://pubmed.ncbi.nlm.nih.gov/17573981/)]
77. Cro S, Morris TP, Kenward MG, Carpenter JR. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: A practical guide. *Stat Med* 2020 Sep 20;39(21):2815-2842. [doi: [10.1002/sim.8569](https://doi.org/10.1002/sim.8569)] [Medline: [32419182](https://pubmed.ncbi.nlm.nih.gov/32419182/)]
78. Rubin DB. Inference and Missing Data. *Biometrika* 1976 Dec;63(3):581. [doi: [10.2307/2335739](https://doi.org/10.2307/2335739)]
79. E9(R1) Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. ICH. URL: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf [accessed 2021-12-05]
80. Tan P, Cro S, Van Vogt E, Szigeti M, Cornelius VR. A review of the use of controlled multiple imputation in randomised controlled trials with missing outcome data. *BMC Med Res Methodol* 2021 Apr 15;21(1):72 [FREE Full text] [doi: [10.1186/s12874-021-01261-6](https://doi.org/10.1186/s12874-021-01261-6)] [Medline: [33858355](https://pubmed.ncbi.nlm.nih.gov/33858355/)]
81. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *The Journal of Machine Learning Research* 2003;3:993-1022. [doi: [10.5555/944919.944937](https://doi.org/10.5555/944919.944937)]
82. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010 Mar 23;340(jul21 1):c869-b2535 [FREE Full text] [doi: [10.1136/bmj.c869](https://doi.org/10.1136/bmj.c869)] [Medline: [20332511](https://pubmed.ncbi.nlm.nih.gov/20332511/)]
83. Gregory RJ, Schwer Canning S, Lee TW, Wise JC. Cognitive Bibliotherapy for Depression: A Meta-Analysis. *Professional Psychology: Research and Practice* 2004 Jun;35(3):275-280. [doi: [10.1037/0735-7028.35.3.275](https://doi.org/10.1037/0735-7028.35.3.275)]
84. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry* 2021 Oct;20(3):318-335 [FREE Full text] [doi: [10.1002/wps.20883](https://doi.org/10.1002/wps.20883)] [Medline: [34505369](https://pubmed.ncbi.nlm.nih.gov/34505369/)]

85. de Gennaro M, Krumhuber EG, Lucas G. Effectiveness of an Empathic Chatbot in Combating Adverse Effects of Social Exclusion on Mood. *Front Psychol* 2019;10:3061 [FREE Full text] [doi: [10.3389/fpsyg.2019.03061](https://doi.org/10.3389/fpsyg.2019.03061)] [Medline: [32038415](https://pubmed.ncbi.nlm.nih.gov/32038415/)]
86. Clarke J, Proudfoot J, Whitton A, Birch M, Boyd M, Parker G, et al. Therapeutic Alliance With a Fully Automated Mobile Phone and Web-Based Intervention: Secondary Analysis of a Randomized Controlled Trial. *JMIR Ment Health* 2016 Feb 25;3(1):e10 [FREE Full text] [doi: [10.2196/mental.4656](https://doi.org/10.2196/mental.4656)] [Medline: [26917096](https://pubmed.ncbi.nlm.nih.gov/26917096/)]
87. Casas J, Spring T, Daher K, Mugellini E, Abou KO. Enhancing Conversational Agents with Empathic Abilities. In: IVA '21: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents. 2021 Presented at: 21st ACM International Conference on Intelligent Virtual Agents; September 14-17, 2021; Virtual Event Japan. [doi: [10.1145/3472306.3478344](https://doi.org/10.1145/3472306.3478344)]
88. Morris RR, Kouddous K, Kshirsagar R, Schueller SM. Towards an Artificially Empathic Conversational Agent for Mental Health Applications: System Design and User Perceptions. *J Med Internet Res* 2018 Jun 26;20(6):e10148 [FREE Full text] [doi: [10.2196/10148](https://doi.org/10.2196/10148)] [Medline: [29945856](https://pubmed.ncbi.nlm.nih.gov/29945856/)]
89. Kocaballi AB, Berkovsky S, Quiroz JC, Laranjo L, Tong HL, Rezzadegan D, et al. The Personalization of Conversational Agents in Health Care: Systematic Review. *J Med Internet Res* 2019 Nov 07;21(11):e15360 [FREE Full text] [doi: [10.2196/15360](https://doi.org/10.2196/15360)] [Medline: [31697237](https://pubmed.ncbi.nlm.nih.gov/31697237/)]
90. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell* 2019 Sep 02;1(9):389-399. [doi: [10.1038/s42256-019-0088-2](https://doi.org/10.1038/s42256-019-0088-2)]

Abbreviations

AI: artificial intelligence
ANCOVA: analysis of covariance
AS: Acceptability Scale
CBT: cognitive behavioral therapy
CSMHSS: College Students Mental Health Screening Scale
ICBT: internet-delivered cognitive behavioral therapy
ITT: intention-to-treat
LDA: Latent Dirichlet Allocation
MAR: missing at random
MCAR: missing completely at random
MI: multiple imputation
MNAR: missing not at random
NLP: natural language processing
PHQ-9: 9-item Patient Health Questionnaire
PP: per-protocol
UMUX-LITE: Usability Metric for User Experience-LITE
VR: virtual reality
WAQ: Working Alliance Questionnaire

Edited by T Leung, V Arnold, H Gouda; submitted 02.07.22; peer-reviewed by E Bunge, I Ezawa; comments to author 23.09.22; revised version received 14.10.22; accepted 03.11.22; published 21.11.22

Please cite as:

He Y, Yang L, Zhu X, Wu B, Zhang S, Qian C, Tian T

Mental Health Chatbot for Young Adults With Depressive Symptoms During the COVID-19 Pandemic: Single-Blind, Three-Arm Randomized Controlled Trial

J Med Internet Res 2022;24(11):e40719

URL: <https://www.jmir.org/2022/11/e40719>

doi: [10.2196/40719](https://doi.org/10.2196/40719)

PMID: [36355633](https://pubmed.ncbi.nlm.nih.gov/36355633/)

©Yuhao He, Li Yang, Xiaokun Zhu, Bin Wu, Shuo Zhang, Chunlian Qian, Tian Tian. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org/>), 21.11.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.