Computers in Human Behavior: Artificial Humans 2 (2024) 100081

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans

User engagement, attitudes, and the effectiveness of chatbots as a mental health intervention: A systematic review

Sucharat Limpanopparat, Erin Gibson, Dr Andrew Harris *

NTU Psychology, Nottingham Trent University, 50 Shakespeare Street, Nottingham, NG1 4FQ, UK

ARTICLE INFO

Keywords:
Chatbot
Mental health intervention
Engagement
User attitude
Effectiveness
Systematic review

ABSTRACT

Background: In recent years, chatbots developed for mental health intervention purposes have been widely implemented to solve the challenges of workforce shortage and accessibility issues faced by traditional health services. Nevertheless, research assessing the technologies' potential and risks remains sporadic.
Purpose: This review aims to synthesise the existing research on engagement, user attitude, and effectiveness of psychological chatbot interventions.
Method: A systematic review was conducted using relevant peer-reviewed literature since 2010. These studies were derived from six databases: PubMed, *PsycINFO*, *Web of Science*, *Science Direct*, *Scopus* and *IEEE Xplore*.
Results: Engagement level with chatbots that complied with digital intervention standards, lead to positive mental health outcomes. Although users had some uncertainties about the usability of these tools, positive attitudes towards chatbots regarding user experience and acceptability were frequently identified due to the chatbots' psychological capabilities and unique functions. High levels of outcome efficacy were found for those with depression. The differences in demographics, psychological approaches, and featured technologies could also influence the extent of mental health chatbot performances.
Conclusion: Positive attitudes and engagement with chatbots, as well as positive mental health outcomes, shows chatbot technology is a promising modality for mental health intervention. However, implementing them amongst some demographics or with novel features should be carefully considered. Further research using mainstream mental health chatbots and evaluating them simultaneously with standardised measures of engagement, user attitude, and effectiveness is necessary for intervention development.

1. Introduction

The global demand for superior mental health care is becoming more urgent and challenging due to a lack of human-based resources and the COVID-19 crisis in recent years (Oladeji & Gureje, 2016; Smith et al., 2020). Chatbot technology (digital tools existing in various modalities with the ability to imitate anthropomorphic behaviours and provide task-oriented or emotionally-oriented conversations; Vaidyam et al., 2019) is being implemented to fill this demand as it could provide better availability and cost-effectiveness than the human-operated services (Campion et al., 2022). Its popularity can be evidenced by the increasing number of mental health apps publicly launched, estimated at around 10,000–20,000 (Clay, 2021), with chatbot-generated conversation being one of the most common categories besides mediation and journalling (Wasil et al., 2021). With these significant investments and interests, the international market for healthcare chatbots was predicted to be worth 250.9 million dollars in 2022 and will increase by 19.8% by 2028 (IMARC Group, 2022).

In contrast with their popularity, quality assessment of chatbots for mental health benefits is still in the infancy phase but is required to assure safety and effectiveness. For example, Lau et al. (2020) found that only 2% of over a thousand mental health applications people can easily download from stores have been empirically supported. Based on the researchers' observation, a limited number of academic reviews put effort into critically accumulating the test results of these chatbots (e.g., Abd-Alrazaq et al., 2020, 2021; Gaffney et al., 2019; Griffin et al., 2020; Lim et al., 2022), finding overall somewhat positive trends but still struggling to create congruent conclusions in several aspects due to the diversity of chatbot characteristics and sample populations. For instance, the primary function of chatbots can be various, from minimal-scale emotional and stress management for general persons to structured counselling for clinical populations of different age groups (e.g., Bennion

* Corresponding author.
E-mail address: andrew.harris02@ntu.ac.uk (D.A. Harris).

https://doi.org/10.1016/j.chbah.2024.100081
Received 24 October 2023; Received in revised form 18 June 2024; Accepted 1 July 2024
Available online 2 July 2024

et al., 2020; Iglesias et al., 2022; Ludin et al., 2022). The therapeutic approach behind each chatbot is either cognitive-behavioural therapy (CBT), which has been well-knowingly accepted for its effectiveness in face-to-face and digital mental health interventions (DMHIs) (Andrews et al., 2018; David et al., 2018) or other methods, including a combination of approaches. Additionally, the technical approach in the chatbots for comprehending and generating interactions is widely developed, ranging from predefined scripts by rule-based technique to complex language processing by Artificial intelligence (AI). These differences in the chatbot features and user demographic could facilitate or obstruct chatbot usage (Boucher et al., 2021), as found in overall DMHIs (Borghouts et al., 2021; Wu et al., 2021). Nevertheless, no review explicitly investigated them in the chatbot focus.

According to the New Medical Research Council guidance (Craig et al., 2008), to conduct efficient evaluation, the researchers should orderly test the feasibility and retention, acceptability, and effectiveness during the development of their health innovations. This suggestion could be supported by the Technology Acceptance Model (Davis, 1989) and the therapeutic alliance model (Ardito & Rabellino, 2011), highlighting the links between these factors. Taken together, they proposed that the perceptions of users (e.g., perceived relationship, usefulness, acceptance) toward the new digital intervention can determine the extent of users' interaction with it, leading to the success rate of its psychological outcome. Thus, a comprehensive review of the existing evidence on engagement, user attitude, and effectiveness of chatbots for mental health advantages is required.

1.1. Mental health chatbots and engagement

Mental health interventions are most likely effective when sustained adherence as recommended is achieved (Fleming et al., 2018). Nevertheless, digital-mediated treatments have been criticised for their poor engagement, representing a high dropout rate and low usage time (Donkin et al., 2011; Fleming et al., 2018). One possible explanation for adherence issues in these 'self-guided' tools may be the lack of response from human staff (Borghouts et al., 2021), which may make users feel lost and less motivated to use them. Although a chatbot is also one type of DMHI, it might overcome this limitation because of its natural and active interactions (e.g., comforting with emphatic language, reflecting on user's stories), similar to those happening in client-therapist conversations. This assumption was supported by the increased engagement with a smoking cessation application when a chatbot was embodied (Perski et al., 2019). However, the engagement-facilitating ability of chatbots is an ongoing gap because little research has been investigated on this topic. In one systematic review, engagement in mental health chatbots appeared highly variable, ranging from very low to high (Gaffney et al., 2019). They suggested that the causes could be the variety in the chatbot's functions and design. However, they could not draw any conclusions because these factors were not explicitly discussed in the original studies. The high level of baseline psychological distress could be another reason for the poor engagement, as traditional intervention studies commonly found a relationship between them (Dixon et al., 2016). Conversely, a chatbot intervention study reported higher engagement in those with more severe issues (Gabrielli et al., 2021). Therefore, it is still unclear whether the initial severity of targeted problems would boost or decline the degree of engagement with mental health chatbots.

1.2. Mental health chatbots and user attitude

Prior evidence suggested that mental health chatbots generally receive acceptance and are reasonably perceived as useable by users in various conditions and intervention types, such as emotional management for the clinical population and promoting/preventive mobile applications for non-clinical target groups (Abd-Alrazaq et al., 2021; Ahmed et al., 2022; Griffin et al., 2020). Positive comments for this technology usually are about its empathic, non-judgmental, and human-likeness manners, along with the beneficial psychology knowledge it provides (Abd-Alrazaq et al., 2021; Ahmed et al., 2022; Vaidyam et al., 2019). These positive experiences could lead to the establishment of a more complex attitude concept, therapeutic alliance, which is the key contributor to successful psychotherapy, as seen in traditional forms of therapy, and many professionals believed that a chatbot-based intervention should also achieve it (Lederman & D'Alfonso, 2021). However, others argued that this bond in the user-chatbot interactions does not necessarily build in the same sense as human-operated interventions since some of DMHIs were effective despite the lack of therapeutic alliance (Ellis-Brush, 2021; Tremain et al., 2020). Despite favourable attitudes, some research participants reported negative experiences and perceptions regarding chatbots as repetitive and, boring, unfunctional, making them feel bothered (Ahmed et al., 2022; Boucher et al., 2021; Inkster et al., 2018). Aside from the technical errors, individual differences could partially be accountable for the variety in user attitudes. For example, older adults were more likely to perceive chatbots as hostile and untrustworthy (Manh-Tung et al., 2023). Since little research has investigated these demographics in mental health chatbot acceptance (Abd-Alrazaq et al., 2019), their possible impacts on chatbot use, which are essential for practical development, remain unclear.

1.3. Effectiveness of mental health chatbots

Even though DMHIs that have passed rigorous tests are usually not the ones widely used by real-world users (Lau et al., 2020), some well-known chatbot interventions, namely Woebot (a CBT-based bot targeting depression), have shown success in depression reduction (Fitzpatrick et al., 2017). Congruently, a few academic reviews in relevant areas (Abd-Alrazaq et al., 2020; Gaffney et al., 2019; Lim et al., 2022) found that chatbots seem to be most effective in dealing with depression both in the clinical and general groups. The results for anxiety, the second most targeted issue, and other problems (e.g., stress, phobias, emotional distress) were conflicting. Chatbot interventions seem to have a minor effect on mental well-being, with no review to date indicating significant positive outcomes. In addition, regarding researchers' observation, many studies that succeeded in establishing the significant effectiveness of their chatbot-based tools were conducted without control groups. With a lack of comparators, this evidence could lead to the overestimated efficacy of chatbots and should be carefully interpreted. Hence, studies with control groups and a broader range of targeted mental issues are crucial to draw sound conclusions about these chatbots' effectiveness.

1.4. The present review

Although psychological chatbot interventions, which are in high interest of health businesses and the public, could fill some gaps in traditional services (e.g., accessibility, workforce, and cost-effectiveness) and somewhat show satisfying outcomes, more multidisciplinary evaluations should be conducted since their potential benefits (or risks) depend on several interconnected performances including engagement, user attitude, and effectiveness. However, according to our database search, all existing reviews solely focused on either one or two of these competencies in experimental research (Abd-Alrazaq et al., 2020, 2021; Gaffney et al., 2019) or included only specific chatbot types and psychopathological problems (Griffin et al., 2020; Lim et al., 2022). Notably, most of them were conducted without considering demographic factors. Their included articles were published before the pandemic, which has noticeably impacted the chatbot advancement and implementation in public health. Therefore, the relationship between these outcome domains of up-to-date mental health chatbot research remains underexplored.

This review aims to synthesise the current evidence of mental health chatbots' potential in three performance areas: i) Engagement:

adherence or frequency of chatbot usage, (ii) User attitude: emotional and cognitive perceptions toward the experience of using chatbots, and (iii) Effectiveness: mental health outcomes after using chatbots. Combining the broader focus on these interplayed aspects with recent studies would help update and extend the preliminary findings about the potential or lack of mental health chatbots necessary for future research and practice implementation in this rapidly changing industry.

2. Method

A systematic review of the published literature followed PRISMA 2020; Page et al., 2021) to achieve the objectives. The references management software (RefWorks) was used for duplicate removal, and the research collaboration web-based tool (Rayyan) was used for study screening and selection. Regarding the guidance from Hong et al. (2017), extracted data were synthesised using a convergent narrative method to integrate qualitative and quantitative findings efficiently.

2.1. Eligibility criteria

The included studies must: (i) primarily focus on the evaluation of chatbots for mental health purposes with outcomes related to at least one of three aspects (engagement, user attitude, and effectiveness), (ii) use an autonomous chatbot without "The Wizard of Oz" method (in which computer responses are generated by a disguised human operator) (McTear et al., 2016), (iii) use primary and empirical methods including quantitative, qualitative, or mix-methods of any study design, (iv) in case of effectiveness assessment, have at least one comparator (e.g., waitlist control group), (v) be conducted within the end-user groups whether clinical population or non-clinical population, (vi) be written in English, and (vii) be published in peer-reviewed journals from 2010 to the present, as chatbot technology became sufficiently advanced and the public became acclimated with it through personal assistants (such as Siri in IOS operation) at that time (Rapp et al., 2021). The exclusion criteria were: (i) the chatbot function was psychological screening without any therapeutic communication, (ii) the chatbot was a small, embodied part of other digital interventions such as robotics or virtual reality, and (iii) The full paper version cannot be accessed.

2.2. Sources

A systematic search was performed on six electronic databases: PubMed, PsycINFO, Web of Science, Science Direct, Scopus and IEEE Xplore. Studies were collected between January and May 2023.

2.3. Search strategy

Two search term groups were combined. The first group relating to chatbot technology was adapted from previous reviews (Abd-Alrazaq et al., 2019; Ng et al., 2019; Vaidyam et al., 2019), while the other group addressing the psychiatric conditions was obtained from the Medical Subject Headings (MeSH) index. Consequently, the following search terms were applied: ("conversational agent*" OR "conversational bot*" OR "conversational system*" OR "conversational interface*" OR "chatbot*" OR "chat bot*" OR "chat-bot*" OR "smartbot*" OR "smart bot*" OR "smart-bot*" OR "virtual coach*" OR "virtual agent*" OR "embodied agent*" OR "relational agent*") AND ("mental disorder*" OR "mental health" OR "mood disorder*" OR "affective disorder*" OR "mental illness*" OR "psychotic disorder*" OR "obsessive-compulsive disorder*" OR autism OR autistic OR adhd OR "Attention deficit hyperactivity disorder*" OR dementia OR "panic disorder*" OR "phobic disorder*" OR "post-traumatic stress disorder*" OR trauma* OR "anxiety disorder*" OR anxious OR stress* OR phobia OR depress* OR melancholia OR bipolar OR schizophrenia OR psychosis OR "eating disorder*" OR anorexia OR binge-eating OR bulimia OR "drug dependence" OR "substance dependence" OR addiction OR "drug abuse" OR "substance abuse" OR "sleep disorder*" OR psychological OR emotional OR well-being OR wellbeing). Search filters (i.e., published between the years 2010 and 2023, English, peer-reviewed) were used to limit the number of results. The search revealed 3115 articles, of which 1349 were duplicates. Hence, there were an overall of 1766 records considered for inclusion. The search was performed.

2.4. Data collection process

Among the original 1766 articles, 1637 were excluded because their titles and abstract did not match the eligibility criteria. The remaining 130 full-text records were screened, and 97 not meeting the criteria were removed. Therefore, 33 articles were passed to the quality appraisal process. The Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018), validated for systematic reviews of mixed studies, was applied to this process. Under the consensus of two reviewers using the MMAT, two studies were excluded because they failed the second screening question, representing data collection issues: inadequate engagement time for effectiveness assessment (Kamita et al., 2019) and incongruence between research design and type of collected data (Gabrielli et al., 2020). In total, 31 studies with various research quality levels (see Appendix 1) were included in the final stage. See Fig. 1 for a detailed diagram of the process.

3. Results

Apart from the study characteristics (see Appendix 1.), findings from 31 included studies were synthesised and grouped into three categories, based on the outcomes of the study: engagement ($n = 21$), user attitude ($n = 21$), and effectiveness ($n = 19$). Indeed, the number of studies in each category does not add up to the total, as most of the included studies discuss multiple outcomes. A summary of each study's findings can be seen in Table 1.

3.1. Engagement

3.1.1. Attrition rate

Of 14 studies reporting attrition rate, four reported that higher than 50% of all participants, ranging from 72% to 60%, dropped out before completing the procedure. Lavelle et al. (2022) discovered a high rate of 72% attrition, with 11% withdrawing immediately after a chatbot failure (e.g., expressing a loop of repetitive messages and being unable to comprehend participants' messages). Moreover, those who dropped out had more significant positive effects and fewer negative thought beliefs than completers at baseline. Two RCT studies noticed the same attrition rate at approximately 60% in the chatbot and waitlist groups (Hunt et al., 2021) or the self-psychoeducation group (Klos et al., 2021). In addition, Hunt et al. (2021) found that participants lost to follow-up were less stressed and catastrophised less at posttreatment than those staying. Dosovitsky et al. (2020) also found an average 60% withdrawal rate in each chatbot's therapy module. They proposed that the completion of each module was connected to its required time length since shorter modules were more likely to be completed.

In contrast, fewer than 50% of participants in ten studies (ranging from 43.5% to 0%) withdrew before finishing chatbot interventions. Among eight RCT studies, five studies did not find a significantly different number of those incomplete in the assigned intervention between the chatbot group and the control group, with an overall attrition rate of 37.3% (Fitzsimmons-Craft et al., 2021), 24.10% (Liu et al., 2022), 20% (Loveys et al., 2021), 19.6 % (Jang et al., 2021), and 7.8% (Bennion et al., 2020). In addition, Loveys et al. (2021) reported that all six participants who withdrew from the study were elderly; half of their reasons were technical difficulties. Conversely, two RCT studies found lower attrition in the psychological chatbot condition compared with the control conditions (Fitzpatrick et al., 2017; He et al., 2022). One study had 0% attrition, as all 74 participants remained from baseline to

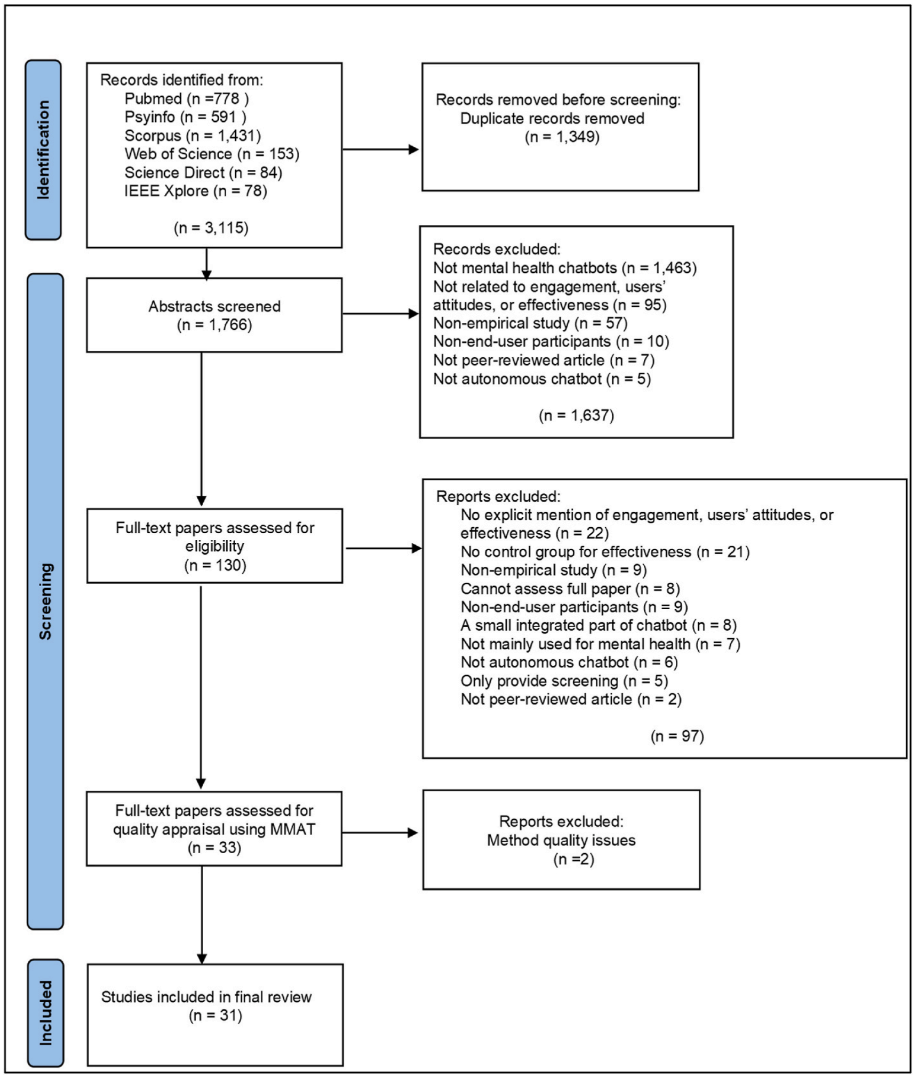

Fig. 1. PRISMA flow diagram demonstrating the process of data selection.

posttreatment timepoint (Fulmer et al., 2018). In a quantitative study by Iglesias et al. (2022), 11.6% of participants downloaded the application but did not engage with any of its modules. Another qualitative study using adolescent participants reported that 43.5% could not complete the online procedure (Dosovitsky & Bunge, 2023).

3.1.2. *Frequency and patterns of use*

The chatbot usage of participants reported in 16 studies varied in measure units (e.g., session, minute, messages) and length of study time.

Two studies collected chatbot usage data within one week. Ludin et al. (2022) found that young users engaged with the chatbot for an average of 11 min. However, participants in another study spent an average of 138 min beyond the research requirements (Loveys et al., 2021).

Among four studies observing the interventions over two weeks, Fitzpatrick et al. (2017) and Ly et al. (2017) reported that users interacted with the chatbots 12.14 and 17.71 times, respectively. Another RCT study indicated that users significantly spent more extended time with the chatbot using the method of level therapy than chatbot implementing humanistic counselling (Bennion et al., 2020). Furthermore, Fulmer et al. (2018) compared the average exchanged messages in the chatbot between group 1 (received daily app reminders and 2-week app access) and group 2 (received biweekly app reminders and 4-week app access). They found comparable proportions of interactions in these two groups.

Of four studies using a 4-week length, Two RCT studies showed a positive association between the amount of engagement (with a mean of 73.8 and 75 min, respectively) and intervention outcomes (Greer et al., 2019; Jang et al., 2021). Greer et al. (2019) found that those using the chatbot had lower anxiety when engaging in more sessions. Similarly, Jang et al. (2021) indicated that ADHD scores were lower when the chatbot was highly used. However, Oh et al. (2020) found no significant relationship between the amount of chatbot usage and panic symptom improvement. In addition, Nicol et al. (2022) reported app use by participants with a mean of 6 days during the same time.

Among four studies with an 8-week period, one study (Hunt et al., 2021) reported that intervention dosage (i.e., the number of modules completed) was slightly related to superior psychological well-being and lower depression score. Martínez-Miranda et al. (2019) showed an average participant's engagement was 84.51 min, while Klos et al. (2021) revealed an average of 472 messages between users and the chatbot. Additionally, Sinha et al. (2022) found that the frequency of app check-in was significantly related to the longer retention period.

Lastly, three studies evaluated engagement for up to six months. Fitzsimmons-Craft et al. (2021) found an average of 16 active days among all participants allocated to the chatbot condition. Two other quantitative studies reported an average engagement of 46 active days over 14 months (Dosovitsky et al., 2020) and 35 active days over 15 months (Iglesias et al., 2022). Furthermore, Iglesias et al. (2022) found

Table 1
A Summary of reviewed studies.

Authors (year)	Types of outcomes	Aims	Sample (*n* at data analysis)	Time length; recommended interactions	Main findings
Lavelle et al. (2022)	Engagement, Effectiveness	To compare the efficacy of two different psychological chatbots (cognitive restructuring vs cognitive defusion) and the non-received chatbot group to decrease negative self-referential thoughts	General adults recruited from university research pools and social media ($n = 62$; 12 males, 49 females, 1 nonbinary); aged 18–68 ($M = 25.98$, $SD = 8.647$)	Baseline, 5 days; at least 5 conversations	Non-completers (with 72% attrition) reported fewer symptoms and higher positive emotions than completers at baseline. No statistically significant differences were observed among the groups.
Zhu, Janssen, Wang, and Liu (2021)	Attitude	To examine the contributing factors that affect the experience and satisfaction of those who use a psychological chatbot during the COVID-19 pandemic	Chinese citizens from 136 families In two cities: Wuhan and Chongqing, recruited from community health centres and social media ($n = 295$; 158 males, 137 females); aged 18+	One time after seven days of use; freely access	Four factors (personalisation, enjoyment, learning, and condition), except voice interaction, were positively associated with user experience and satisfaction when interacting with the chatbot.
Fitzpatrick et al. (2017)	Engagement, Attitude, Effectiveness	To compare the initial performance of an autonomy anxiety and depression programme provided by a chatbot to a self-psychoeducation group.	Universities students in the US who self-reported depression and anxiety, recruited from social media ($n = 58$; 11 males, 47 females); aged 18–28 ($M = 22.2$, $SD = 2.33$)	Baseline, 14 days; daily access	There was a lower attrition rate in the chatbot group. Process facets of the chatbot dominantly affect both positive and negative user experiences more than content and technical components. Those in the chatbot group had more significant improvement in depression, but both groups had the same rate of reduced anxiety.
He et al. (2022)	Engagement, Attitude, Effectiveness	To assess the performance of a CBT-based chatbot for young persons with depression risks during the COVID-19 outbreak, compared with a typical companion chatbot and reading books groups	Chinese university students with an average score of depression recruited from the university counselling centre and social media (*n* 148; 93 males, 55 females); aged 17–21; ($M = 18.78$, $SD = 0.89$)	Baseline, 7 days, one month; daily access	There was less attrition in the mental health chatbot than in other control groups. The mental health chatbot showed higher working alliance and acceptability, whereas no significant intergroup difference in usability. Participants using the mental health chatbot reported the most significant depression reduction over the study period.
Medeiros, Bosse, and Gerritsen (2022)	Effectiveness	To compare whether users know that the interaction provided by the chatbot (perceiving support as Chatbot-generated VS perceiving support as human-generated VS does not receive active support) impacts its effectiveness on levitating daily-life stress	General adults recruited from a crowdsourcing platform ($n = 84$; 12 males, 49 females); aged 20-64	Before and after using the chatbot for three days; daily access	There was a group difference with those perceiving the chatbot as a human having a higher positive feeling than other groups after the conversations. In contrast, the perceiving as chatbot group did not report increased positive feelings differently than the non-active control group.
Dosovitsky et al. (2020)	Engagement	To explore the user's behaviour toward engagement with a chatbot aimed at depression (Tess) to improve the design of it	Users who downloaded Tess from the online store and engaged with at least one of the 12 depression modules on it ($n = 354$)	Over 14 months; Freely access	On average, users participated with the app for 25 min within 46 days of study time. Across 12 modules, the mean completion rate was only 40%, and shorter modules were more likely to be completed than the longer ones.
Philip et al. (2022)	Effectiveness	To test the effectiveness of a chatbot app for insomnia issues management, compared with another sleep diary app	French users who voluntarily downloaded either application ($n = 1024$; 342 males, 682 females); aged 18+	Baseline, 7 days, 17 days ; daily access	The chatbot significantly reduced insomnia complaints and estimated nocturnal sleep patterns in individuals with moderate sleep problems than the diary app. The chatbot group showed significant improvement in users with severe sleep issues, while the sleep diary group did not.
Hunt et al. (2021)	Engagement, Effectiveness	To determine the potential of the newly developed chatbot specifically providing CBT for anxiety among persons having irritable bowel syndrome (IBS), compared with the waitlist group	Patients who had been diagnosed with irritable bowel syndrome recruited via IBS-specific social media sites ($n = 59$); aged 18–63 ($M = 32$, $SD = 10.2$)	Baseline, 8 weeks, 3 months; daily access for 8 weeks	With an overall 60% attrition, Participants who dropped at the follow-up were more likely to be less stressed and catastrophised. Various types of IBS-related anxieties were more significantly decreased in the intervention group over three months. Depression only reduced significantly in the chatbot group in the completer analysis. A higher chatbot engagement was slightly associated with better psychological well-being and less depression.

(*continued on next page*)

Table 1 (*continued*)

Authors (year)	Types of outcomes	Aims	Sample (*n* at data analysis)	Time length; recommended interactions	Main findings
Ta et al. (2020)	Attitude	To explore the kinds of psychological supports that can be emerged when individuals interact with a companion chatbot	Users who downloaded the chatbot (Replika): Study 1 using application reviews app market platform (*n* = 1854 reviews) study 2 using an online survey (*n* = 66; 36 males, 30 females); aged 17–68 (*M* = 32.64, *SD* = 13.89)	N/A	There were four types of social support (companionship, emotional, appraisal, And informational) found in both study 1 and 2. The negative experiences about the uncanny valley and non-sense computer-generated responses were only noticed in study 1.
Fitzsimmons-Craft et al. (2021)	Engagement, Effectiveness	To evaluate whether a developed chatbot can reduce ED-related psychopathology in women with eating disorders concerns population, compared to the waitlist group	Adult females who had high eating disorders (ED) scores but did not meet the clinical cut-off, recruited via social media and national ED association (*n* = 439); aged 18–30 (*M* = 21.08, *SD* = 3.09)	Baseline, 3 months, 6 months; at least 2 conversations per week	Overall completion was 62.7%, with an average of 4 interactions and 11 min/interaction. The chatbot group reported more remarkable improvement in weight/shape concerns over six months and ED psychological issues over three months, compared with the control group.
Iglesias et al. (2022)	Engagement	To examine the usage pattern and user demographic differences of a chatbot specially designed for those in a Workers' Compensation (WC) with mental health problems	Employees who were unable to work due to a medical incident and had psychological risks were recruited via WC association pools (*n* = 190; 96 males, 94 females); aged 17-75	Over 15 months; freely access	Participants who were female and had higher psychosocial risk factors exhibited a greater onboarding rate, retention rate, and engagement rate than male and lower-risk participants. Additionally, those with more severe injuries demonstrated a higher rate of chatbot usage retention.
Liu et al. (2022)	Engagement, Attitude, Effectiveness	To assess whether a chatbot can be a more effective self-help depression intervention than self-psychoeducation via an ebook	University students who had high depression scores but were not diagnosed, recruited via social media (*n* = 73); aged 19–28 (*M* = 23.08, *SD* = 1.76)	Baseline, 4 weeks, 8 weeks,12 weeks,16 weeks; freely access	There was an overall 24.10% attrition rate. The process aspects of the chatbot affect user experience the most, and some users felt that it enhanced their willingness to receive professional services. The chatbot evidentially increased therapeutic alliance and reduced depressive and anxious symptoms more than reading the psychoeducational book.
Ludin et al. (2022)	Engagement, Attitude	To estimate an anxiety management chatbot's adherence and usability level for young populations during the pandemic.	Individuals who were in age 13–24 (*n* = 127; 37 males, 90 females)	5 days after using it; freely access	After engaging for an average of 11 min, the chatbot was perceived as accessible, useable, and necessary among young users. Based on user feedback, more comprehensive content about living problems and better conversation dynamics were most required.
Oh et al. (2020)	Engagement, Attitude, Effectiveness	To evaluate user perceptions and psychological outcomes from using a developed chatbot to deal with panic issues, compared with self-psychoeducation	Korean patients suffering from mild to severe panic symptoms who were sourced from an outpatient clinic (*n* = 41; 20 males, 21 females); aged 19-60	Baseline, 4 weeks; freely access	The amount of chatbot usage did not link to the degree of panic disorder reduction. Process elements were the primary mentioned aspect of user experience evaluation. The chatbot was not rated on usability differently from the psychoeducation method. There were between-group differences that significant panic severity, social phobia, and helplessness improvement only were noticed in the chatbot group. There were no significant differences in general depression or anxiety scores.
Nicol et al. (2022)	Engagement, Attitude, Effectiveness	To investigate the preliminary user attitudes and efficacy of a stan-alone chatbot intervention for adolescents suffering from moderate depression, compared with the waitlist group	Adolescents who were newly diagnosed with moderate depression and anxiety, recruited from a psychiatric centre (*n* = 17; 2 males, 15 females); aged 13-17	Baseline, 2 weeks, 4 weeks; freely access	On average, young users used the chatbot for six active days, rated it as highly useable and acceptable, and made them feel positive after using it. The decreased depressive symptom was more prominent in those participating with the chatbot than in the control condition.
Bennion et al. (2020)	Engagement, Attitude, Effectiveness	To compare the extent of adherence, usability and effectiveness of two different	Old adults who had no clinical psychiatric disorder but experienced a problem causing	Baseline, posttreatment, 2 weeks;	Participants significantly interacted with the Method of Levels (MOL) therapy chatbot longer than the

(*continued on next page*)

Table 1 (*continued*)

Authors (year)	Types of outcomes	Aims	Sample (*n* at data analysis)	Time length; recommended interactions	Main findings
		psychological methods embodied in chatbots (method of levels VS humanistic counselling) for decreasing problem-related distress	emotional distress, recruited from university research pools (*n* = 94; 25 males, 69 females); aged 50–91 (*M* = 68.4, *SD* = 6.49)	One-time access with at least 20 min	humanistic counselling chatbot. Although no usability difference was found between the two chatbots, usability could predict the level of perceived helpfulness and reuse intention. The MOL group significantly had a higher reuse intention. Only at two weeks did the MOL group significantly improve in psychological distress than the humanistic group. However, the depression, anxiety, and stress level were not found to be different between the compared chatbots.
Dosovitsky and Bunge (2022)	Engagement, Attitude	To investigate the overall experience and perceptions of adolescents who used a mental health chatbot aimed at depression risks	General us adolescents contacted via parents (*n* = 13); aged 13–18 (*M* = 14.95, SD = 1.49)	One time after a 30 min access	A 43.5% attrition rate was found. Most user feedback was positive, and they rated the possibility of recommending the chatbot to peers at above 6/10. However, Refinements about typing and emojis issues were the most mentioned.
Park et al. (2019)	Attitude	To evaluate the user experience of graduate students when having a motivational interview performed by a chatbot aiming at stress management for future design development	Graduate students who had high stress regarding education life, recruited from social media (*n* = 30; 15 males, 15 females); aged 24 -32	One time after a 30 min access	The chatbot's most frequently reported positive aspects were its ability to discuss the idea of change, facilitated by evocative questions and emphatic responses related to school life struggles. However, based on participant opinions, the chatbot still needed to provide more beneficial information and in-depth and contextualised response.
Stara et al. (2021)	Atititude	To assess whether demographic differences of individuals with dementia could associate with their attitude towards the chatbot exclusively developed as daily life assistance for people having cognitive impairments	Elderly persons who diagnosed with dementia in three countries (Italy, Luxembourg, and the Netherlands) with mild to moderate cognitive impairments (*n* = 55; 20 males, 35 females); aged 60+ (*M* = 70.8, *SD* = 9.9)	Baseline, 4 weeks; freely access	Those who were older, had lower cognitive scores, and had zero experience with tablets rated the usability significantly lower than the younger, had higher cognitive scores, and had used tablets before. Most of the users gained litter to no connection with the chatbot.
Greer et al. (2019)	Engagement, Attitude, Effectiveness	To examine the engagement, user perception, and effectiveness of a chatbot providing positive psychology skills for dealing with depression and anxiety in young adults having cancer, compared with the waitlist group	Youth who had finished cancer therapy within the past five years, obtained from medical facilities (*n* = 45; 9 males, 36 females); aged 18–29 (*M* = 25, *SD* = 2.9)	Baseline, 2 weeks, 4 weeks; freely access	Among those using the chatbot, the amount of engagement positively related to the degree of anxiety improvement. With an average of 6.9/10 recommending others to use the chatbot, the nonjudgmental context and positive content were the dominant reasons for the recommendation. In contrast, perceived uselessness and annoyance were the explanations for not recommending using the chatbot. There was a more excellent improvement in anxiety but not depression in those using the chatbot than those not receiving the intervention.
Suganuma et al. (2018)	Effectiveness	To estimate the primary efficacy of a newly built AI-based chatbot aimed at mental-wellbeing of Japanese individuals, compared with a waitlist control group	Japanese college students, employees and homemakers from an online research company pool who were interested in using the app (*n* = 427; 136 males, 291 females); aged 18+ (*M* = 38.04, *SD* = 10.75)	Baseline, 1 month; daily access with at least 15 days	Those who used the chatbot decreased their psychological distress and increased psychological well-being significantly, while the waitlist group's negative and positive mental factors remained the same.
Fulmer et al. (2018)	Engagement, Attitude, Effectiveness	To compare the overall usage behaviours, user perception and effectiveness of an AI-generated bot with an e-book or delayed access to the chatbot for alleviating depression and anxiety among university students.	US general university students with moderate self-reported depressive and anxious symptoms from 15 universities, recruited from social media (*n* = 74; 21 males, 52 females, one non-identified); aged 18+ (*M* = 22.9)	Baseline, 2 weeks, 4 weeks; freely access	There was almost 0% attrition in all participants. The chatbot received higher levels of satisfaction and acceptability than the e-book. The more prominent symptom improvements of depression and anxiety were found in those assigned to the chatbot group.

(*continued on next page*)

Table 1 (*continued*)

Authors (year)	Types of outcomes	Aims	Sample (*n* at data analysis)	Time length; recommended interactions	Main findings
Martínez-Miranda et al. (2019)	Engagement, Attitude	To assess the engagement pattern and perception of targeted users when using the chatbot embodied with a virtual agent for persons with suicidal risks to assist them in maintaining the healthy mental status	Mexican individuals identified by mental health professionals with any suicidal behaviour (ideation, planning or attempt) but not at the present time (n = 12); aged 20-53	One time after 8 weeks of use; daily access	Over eight weeks, participants interacted with the chatbot for 84.51 min. Approximately 60% of them also accepted the emotional competence of this intervention. Nevertheless, the pleasant appearance of the chatbot could be improved since the average score was near the medium point. According to user opinions, the voice of the virtual avatar in the chatbot should be more natural.
Jang et al. (2021)	Engagement, Attitude, Effectiveness	To estimate the efficiency in terms of adherence, user perception, and clinical outcomes of a chatbot application for ADHD problems, compared with a self-psychoeducation group	Korean patients who reported significant attention problems (n = 37); aged 19-60	Baseline, 4 weeks; daily access	The higher amount of chatbot adherence (with an average of 75 min) was significantly related to lower ADHD symptom scores. Empathised and friendly characteristics of the bot were the best acknowledged positive part, while the limited conversation dynamic was the worst perceived aspect. Participants rated a similar satisfaction degree for the chatbot and e-book. The chatbot significantly improved several ADHD sub-scores, while the e-book did not.
Loveys et al. (2021)	Engagement, Attitude, Effectiveness	To empirically test the potential and usefulness of a chatbot intervention dealing with loneliness and stress when individuals were self-isolated during the pandemic, compared with a waitlist group	Adults and elderly who were being self-isolated due to medical conditions and at risk of COVID-19 infection (n = 24); aged 18–69 and 70+ (M = 68.20, SD = 18.95)	Baseline, 1 week; at least 15 min per day	All users dropping out from the intervention were elderly people. However, remain younger and older groups did not engage with the chatbot differently in terms of time spent on the chatbot, which outnumbered the requirement. The human-likeness of it was appreciated, but some components within this still need to be adjusted. Few users preferred to receive psychological services from humans. Significant stress, psychological distress and well-being improvements were not found in the chatbot or control group.
Leo et al. (2022)	Effectiveness	To compare the clinical benefits, both mental and physical, of a chatbot for depression and anxiety in chronic pain patients with the other two control groups (receiving mental care from professionals, only receiving usual physical care)	Adult patients with clinical musculoskeletal pains and had self-reported depression or anxiety (n = 153; 25 males, 128 females); aged 18–86 (M = 55, SD = 15)	Baseline, 2 months; freely access	Patients in the chatbot condition reported clinically improved depression and anxiety over study time. Moreover, these improvements were similar to those who received counselling from psychologists and significantly more than those who received only orthopaedical treatment and similar to those who received counselling from psychologists.
Park et al. (2022)	Attitude	To evaluate the impact of emotional disclosure in a counselling chatbot on user satisfaction and reuse intention, and whether user emotional disclosure and established intimacy can mediate the relationships between these factors, compared with a chatbot only providing relevant information	US adults recruited from a crowdsourcing platform (n = 348; 190 males, 158 females); aged 21–71 (M = 36.9)	One time after access	The chatbot with the ability of emotional disclosure attained higher user satisfaction and a greater likelihood of reuse than the informational chatbot. Moreover, user emotional disclosure and perceived intimacy were positively correlated with the degree of chatbot emotional disclosure, and these user perceptions could serve as the mediators in the associations.
Ly et al. (2017)	Engagement, Attitude, effectiveness	To investigate the overall efficiency of a chatbot app which provided CBT-based conversations, compared with those on a waitlist, for general mental well-being maintaining	Swedish adults recruited via social media (n = 28; 13 males 15 females); aged 20–49 (M = 26.2, SD = 7.2)	Baseline, 2 weeks; at least 14 conversations	The samples showed high engagement beyond the minimum requirement. Three themes (content, functionalities, and medium) emerged from users' feedback, with the content factor being the most influential contributor to user perception of

(*continued on next page*)

Table 1 (*continued*)

Authors (year)	Types of outcomes	Aims	Sample (*n* at data analysis)	Time length; recommended interactions	Main findings
					the chatbot. Only analysis with data from those who adequately interacted with the chatbot revealed significant stress and psychological well-being improvement in the experimental group than in the control group.
Shah et al. (2022)	Attitude	To develop and empirically test the usability of a chatbot aiming at enhancing treatment motivation and self-efficacy among persons with eating disorders	Individuals who were screened positive for eating disorders and did not start receiving any treatment ($n = 21$; 1 male, 20 females)	One time after 2 weeks of use; Freely access	The average usability score in the final evaluation cycle, at 85.8, of the developed chatbot was higher than the general acceptable point.
Sinha et al. (2022)	Engagement	To assess engagement patterns of patients who were suffering from chronic pains with an AI-based chatbot for alleviating depressive and anxious issues	Adults who went to a specialist centre for chronic pains and reported clinical level of depression or anxiety ($n = 51$)	One time after 8 weeks of use; Freely access	Participants interacted with the chatbot for 33.3 sessions over the eight weeks. 20% of users continually used the chatbot even after the end of the study. Moreover, the amount of morning app check-in was positively associated with the retention period.
Klos et al. (2021)	Engagement, attitude, effectiveness	To evaluate the potential of a mental health chatbot, compared with a psychoeducational book, in terms of adherence, user perceptions, and depression and anxiety improvements in college student populations	General Argentinian college students ($n = 73$; predominantly being females); aged 18-33	At baseline, 8 weeks; freely access	There was a high overall attrition rate of 60%. Participants generally had more positive opinions than neutral or negative ones towards the chatbot intervention. Positive feedback also was positively related to engagement. Although there were no significant between-group differences, only those in the chatbot group showed a decreased trend of both depression and anxiety over eight weeks.

that persons with more psychosocial risks established higher engagement with the chatbot.

3.2. User attitude

3.2.1. User experience

Fourteen studies in the present review investigated user experience that involved all aspects of users' perceptions during the human-computer interaction process (Lallemand et al., 2015).

Of seven studies using the two open-ended questions approach, six studies (Fitzpatrick et al., 2017; Fulmer et al., 2018; He et al., 2022; Jang et al., 2021; Liu et al., 2022; Oh et al., 2020) demonstrated that process aspects (e.g., empathy and friendly personality, accessibility via daily reminders, learning facilitator) outnumbered the content aspects (e.g., helpful knowledge and practical training) in terms of reported positive experience. They also found a similar trend towards negative experiences in which users mentioned process aspects (limited language comprehension, expressing, and boringness) more frequently than content aspects (inadequate novel content, content length, and inappropriate emoticons) and technical errors. The remaining study (Loveys et al., 2021), implementing a chatbot with a virtual agent, found that participants generally favoured its human-like attributes, the delivered content, and the novel feeling of interacting with advanced technology. However, participants suggested that the human-likeness of the chatbot's actions and voice, followed by the personalisation of conversation topics, needed to be improved.

Two other studies categorising participants' feedback into three groups (positive, neutral, and negative) found a higher propensity towards positive feedback, indicating that chatbots were perceived as accessible tools for mental-related boost (Dosovitsky & Bunge, 2023; Klos et al., 2021).

The first of two studies using semi-structured interviews to explore user experience (Park et al., 2019) highlighted the different opinions towards specific components of the embodied psychotherapeutic method. The chatbot was most favoured in the evoking question ability, which helped users to reflect on themselves. However, these computer-generated responses were still not entirely up to their expectations since the chatbot had apparent problems with inadequately helpful information and too general responses. User opinions from Another study (Ly et al., 2017) were divided into three themes, with the content theme (appreciation of learning or life reflection and complaining about the content repetitiveness) being the dominant part.

Furthermore, Ta et al. (2020) conducted two linked studies on the AI-generated companion chatbot, one analysing application reviews on online stores and the other using open-ended questions. Four themes reflecting different types of social support emerged, with companionship support being mentioned the most in both studies due to its ability to alleviate loneliness and grasp complex interpersonal behaviours. Only a few unfavourable experiences complaining about unsettling feelings to its much human likeness and illogical responses were demonstrated in study 1. Additionally, Nicol et al. (2022) assessed immediate user experience by showing various mood-related emojis for them to choose from after each conversation. Most participants selected positive ones. Finally, survey-based research by Zhu et al. (2021) found that personalisation, enjoyment, learning, and condition factors positively affected user experience.

3.2.2. Usability

Twelve studies assessed the usability and user satisfaction of implemented chatbots, which can be defined as the degree to which a tool can achieve specific objectives, leading to satisfaction within the context of usage (International Organization for Standardization, 2013).

Among four studies using the System Usability Scale (SUS; Brooke, 1996), two RCT studies found no significant differences between the psychological chatbot and a psychological book (Oh et al., 2020) or a typical chatbot (Bennion et al., 2020). Another mix-method research reported that older individuals with lower cognitive function and who never used tablets rated SUS scores lower than the younger, higher cognitive function and digitally experienced users (Stara et al., 2021). It is worth mentioning that SUS scores from these three studies (ranged 63.56 to 66.2) were slightly below the cut-off (68 scores) for an acceptable level of intervention usability. The chatbot targeting individuals with eating disorders from Shah et al. (2022) is the only intervention with above-average usability, with 85.8 scores. The feedback from users demonstrated that the chatbot had excellent usability in general but should be improved in too typical content and robotic language. Additionally, Nicol et al. (2022), using their developed scale, reported the high-level usability of the chatbot in young clinical samples.

Four randomised experiments used the same five-point scales to assess overall and content satisfaction. Three of the studies found a higher level of two satisfaction types in samples assigned to the mental health chatbot than those in eBook (Fitzpatrick et al., 2017; Fulmer et al., 2018) or general chatbot (He et al., 2022) groups. Despite the preference for conversational agents, Jang et al. (2021) reported no significant satisfaction difference between the chatbot and psychoeducation groups. Similarly, using a different measure, Liu et al. (2022) showed that both the chatbot and bibliotherapy conditions had similar intervention satisfaction.

In a study evaluating usability based on qualitative information (Ludin et al., 2022), young participants emphasised the need for mental health chatbots due to elevated levels of stress, especially during the COVID-19 pandemic. Additionally, they identified requirements for broader content concerning living difficulties (i.e., financial stress and domestic violence) and more dynamic conversation. Finally, Zhu et al. (2022) found that personalisation, enjoyment, learning, and condition factors positively impacted user satisfaction.

3.2.3. Acceptability

Thirteen studies examined the acceptability of chatbots, which refers briefly to the psychological and emotional acceptance of the therapeutic process and content provided in chatbots, leading to the intention to use them in the future (Davis, 1989; Sekhon et al., 2017).

Four RCT studies used the same rating questions to evaluate the acceptability. Three of them found higher levels of these scores in the chatbot condition (Fitzpatrick et al., 2017; Fulmer et al., 2018; He et al., 2022), while one did not discover any significant difference (Jang et al., 2022).

Moreover, He et al. (2022), who measured the working alliance, noticed a higher degree of it among individuals in the therapeutic chatbot group, in line with another work by Liu et al. (2022). They indicated this alliance was more prominent in those interacting with the AI-operated chatbot than in the self-psychoeducation group.

The other two studies also discussed the acceptability of chatbots in building relationships with users. Loveys et al. (2021) found that participants perceived some degree of a bond between them and the AI-based chatbot and might voluntarily use it again. In contrast, elderly participants with dementia risks from Stara et al. (2021) reported a lower degree of a perceived relationship, with 60% feeling no to little connection with the chatbot. Open-end questions in the same study also indicated that 60% of users got distracted by or did not receive any advantages. At the same time, the rest viewed it as a life assistant or reminder. Unlike the opinions of older participants, 80% of adolescent users from Nicol et al. (2022) found the app acceptable and liked using the chatbot.

Three studies evaluated the extent to which participants would use or recommend the chatbots to others in the future. Bennion et al. (2020) found that participants assigned to the chatbot with the Method of Levels (MOL) therapy were more willing to reuse it than those using the humanistic counselling chatbot. Additionally, users from Dosovitsky and Bunge (2022) rated whether they advocated chatbot use to peers, with a mean score of 6/10, and 64.3% of users indicated that they would use the chatbot again in the future. Congruently, the third study found that participants would recommend the chatbot to friends, with a mean score of 6.9/10 (Greer et al., 2019). Further feedback from responders indicated that the nonjudgmental environment and positive psychological content within the chatbot were the reasons for the recommendation. In contrast, perceived unhelpfulness and annoyance were key reasons for not recommending the technology.

Finally, the remaining two studies focused on the ability of chatbots to portray acceptable emotional behaviours. The first RCT study compared two chatbots, in which one could express its own emotions while the other only gave related information (Park et al., 2022). The results showed that those using the emotional disclosure chatbot had higher reuse intention and satisfaction than those using the factual information chatbot. Furthermore, the extent of the chatbot's emotional disclosure was positively associated with users' emotional disclosure and intimacy level. In the second study (Martínez-Miranda et al., 2019), emotional competence, the ability to deliver coherent and consistent emotional behaviour, of chatbot-based suicidal prevention was evaluated. The chatbot was significantly accepted by almost 60% of the participants. Based on qualitative feedback, most suggestions related to improving the virtual avatar's voice to be less robotic.

3.3. Effectiveness

3.3.1. Depression

Of thirteen studies assessing depressive symptoms, seven of them used Patient Health Questionnaire (PHQ-9; Kroenke et al., 2001). Participants in nonclinical (Fitzpatrick et al., 2017; Fulmer et al., 2018; He et al., 2022; Liu et al., 2022) and clinical populations (Nicol et al., 2022) who used chatbots from five RCT studies showed a similar trend of more remarkable improvement than the comparators, with effect size ranged from moderate to high. Another study that tested the chatbot for anxiety in irritable bowel syndrome (Hunt et al., 2021) also found significantly decreased depression via this scale in the experimental group. However, Klos et al. (2021) was the only study that did not notice any differences between groups using the PHQ-9 measure. In comparison, Fitzsimmons-Craft et al. (2021), using a different version of the measure, PHQ-8, also indicated no statistical differences. However, participants either assigned Eds-management chatbot or waitlist reported a similar significant improvement over time.

Among two studies using the Patient-Reported Outcomes Measurement Information System (PROMIS; Cella et al., 2007) for depression assessment, one study by Leo et al. (2022) found that chronic pain clients who received the chatbot intervention had a significant decrease. This clinical improvement also was comparable to that in the traditional counselling group. Another RCT in a clinical cancer sample revealed that both compared groups showed the same lowered trend of depression with no significant intergroup difference (Greer et al., 2019).

The last three studies measured depressive symptoms with different tools (Bennion et al., 2020; Jang et al., 2021; Oh et al., 2020), and all of them did not notice any group differences towards depression scores.

3.3.2. Anxiety and stress

Among fourteen studies that reported the effectiveness of chatbots on anxiety and stress, five used The Generalized Anxiety Disorder Scale (GAD-7; Spitzer et al., 2006). Two studies found a more significant improvement in anxiety in the chatbot group (Fulmer et al., 2018; Liu et al., 2022). Conversely, the other three found that chatbot users rated significantly lower anxiety over time, but these psychological benefits were not statistically different from those in the control condition (Fitzpatrick et al., 2017; Fitzsimmons-Craft et al., 2021; Klos et al., 2021). Two RCTsevaluating anxiety with different measures (Bennion et al., 2020; Oh et al., 2020) also found only an insignificant difference between groups. Bennion et al., 2020 used The Depression, Anxiety, and Stress Scales 21 (DASS-21; Lovibond & Lovibond, 1995), whereas Oh et al. (2020) used the Panic Disorder Serbvity Scale (PDSS; Shear et al., 1977).

Two clinical-based studies using PROMIS showed a similar trend of significant anxiety reduction in the chatbot condition, and these improvements were more excellent than the outcome from the waitlist group (Greer et al., 2019) or equivalent to those receiving professional services (Leo et al., 2022).

Among four studies focusing on stress outcome, three studies assessing stress via the Perceived Stress Scale (Lee, 2012) failed to observe any intergroup difference (Jang et al., 2021; Loveys et al., 2021; Ly et al., 2017). Work by Medeiros et al. (2022) was the only one measuring it through the Affect Grid-a single-item measure introduced by Russell et al. (1989). They found that all participants in three conditions (interacting with the chatbot but perceiving as human, interacting with the chatbot and perceiving as chatbot, non-interactive control group) had significantly lower stress after the conversation.

Finally, in the randomised trial by Hunt et al. (2021), the chatbot for IBS patients showed significant effectiveness, with individuals in the experimental group reporting greater improvements in IBS-related anxiety symptoms than those in the waitlist group.

3.3.3. Positive and negative emotions

Among four studies that provided the emotional outcomes via the positive affect and negative scales (PANAS; Watson et al., 1988), Fulmer et al. (2018) was the only study indicating significant differences between the chatbot and control groups. The remaining three studies did not observe any significant improvement in either the chatbot or control condition (Fitzpatrick et al., 2017; Lavelle et al., 2022; Liu et al., 2022).

An additional four studies addressed several types of negative emotions. Bennion et al. (2020), who compared two chatbots with different psychological approaches, found that participants in both groups had significantly less problem-related distress over time. This improvement was more remarkable in those using the MOL chatbot. Similarly, non-RCT research by Suganuma et al. (2018) evaluated psychological distress and found a significantly lower score in the chatbot group than in the waitlist group. Greer et al. (2019) found that both patients in the chatbot and control groups appeared to report less negative emotion, but this was not a statistically significant change. The final study (Loveys et al., 2021), also reported no significant difference in negative emotion scores.

3.3.4. Psychological wellbeing

Among three studies that discussed the effect of chatbots on positive mental health, all found no significant improvement in the intervention and waitlist groups (Loveys et al., 2021; Ly et al., 2017). Conversely, the third study using the World Health Organization-Five Well-Being Index demonstrated a significant-but-small improvement in the chatbot group over the control group (Suganuma et al., 2018).

3.3.5. Miscellaneous mental health concerns

Five studies individually addressed various mental health issues, including sleep-related problems (Philip et al., 2022), eating disorders (Fitzsimmons-Craft et al., 2021), self-referential thoughts (Lavelle et al., 2022), panic disorders (Oh et al., 2020), and ADHD (Jang et al., 2021). Four of these studies found significant differences between groups with either more noticeable improvements in the chatbot conditions (Fitzsimmons-Craft et al., 2021; Oh et al., 2020; Philip et al., 2022) or only a significant reduction in those using the chatbot (Jang et al., 2021). However, one RCT study failed to find any between-group difference regards to self-negative thoughts (Lavelle et al., 2022).

4. Discussion

The present study systematically reviewed empirical evidence on three performance domains: engagement, user attitude, and effectiveness of chatbot interventions with different mental health advantages. Two-thirds of the included studies were published within the last three years (2021–2023; see Table 1), indicating that the use of conversational agents for dealing with mental health problems is expanding.

4.1. Engagement with mental health chatbots

The chatbot intervention generally received an adequate level of engagement since 10 of 14 studies revealed attrition rates lower than 50%, which is the recommended cut-off threshold for effective DMHIs (Nwosu et al., 2022). None of the chatbots in the four studies with high attrition (e.g., Hunt et al., 2021; Lavelle et al., 2022) featured a reminder function, but others with low drop-out rates did. Therefore, introducing a usage facilitator could lessen the engagement problem, as supported by the positive relationship between the check-in feature and more prolonged usage behaviour (Sinha et al., 2022). A typical session of classic conversational treatment lasts at least 30 min (Roach et al., 2023), but only five studies in this review demonstrated that their user engagement over the study period could last that long (Fitzsimmons-Craft et al., 2021; Loveys et al., 2021; Greer et al., 2019; Jang et al., 2021; Martnez-Miranda et al., 2019). These findings could signal that consumers frequently interact with chatbots in a brief manner that might be insufficient to create major psychological changes. Contrastingly, participants in a study by Bennion et al. (2020) demonstrated a substantial improvement after less than half an hour of interacting with the intervention and users from two experiments also preferred shorter usage times (Dosovitsky et al., 2020; Dosovitsky & Bunge, 2023). Furthermore, a prior meta-analysis (Lim et al., 2022) suggested that chatbot psychotherapies with fewer than ten sessions improved depression more than their comparators. Taking these into account, the short duration spent on chatbots does not always need to be considered as a negative attribute, and effective ways to engage with this digital tool may differ from what has been advised in the offline ones.

4.2. User attitude towards mental health chatbots

Overall, mental health chatbots appeared to be perceived more positively than negatively by users from various backgrounds regarding user experiences. A similar pattern of more notable appreciations and recommendations for this intervention's process attributes rather than its content and technological components could imply that the process components dominantly affect user attitude and should be prioritised when building chatbot-based psychotherapy. This technology's acceptability was also widely acknowledged and tended to be higher than the compared interventions (e.g., Fitzpatrick et al., 2017; Fulmer et al., 2018; He et al., 2022). Importantly, the reviewed findings suggested that a therapeutic relationship can be established between users and conversational agents in the mental health context (He et al., 2022; Liu et al., 2022; Loveys et al., 2021), and this bond might, in part, result from the emotional-related competency of chatbots (Martínez-Miranda et al., 2019; Park et al., 2022) which is a pivotal contributor to the therapeutic relationship in a traditional context. However, because chatbots are still incapable of imitating empathy and emotion as therapists, these advantageous partnerships could be boosted by other components. For example, Ly et al. (2017) participants mentioned that chatbots' ability to provide daily interactions made them emotionally attached to and perceive it as part of their routine. Nevertheless, additional explorations are required to establish empirical conclusions. Usability concerns were the most prominent of the three terms under the user attitude area. Three out of four studies reported the usability score, which is appropriate for determining the usability of internet-based therapies (Mol et al., 2020), as below the acceptable point. However, low usability commonly happens when a novel tool is introduced to users due to unfamiliarity and its immature ability (Liang et al., 2018). The only chatbot in this review obtaining a high usability score was tested by combining in-person and remote designs (Shah et al., 2022). Therefore, comprehensive evaluation could lead to greater usability of the chatbot. Furthermore, in their recent conference proceedings (Boyd et al., 2022) highlight the potential importance of usability of chatbots in attitude formation.

4.3. Effectiveness of mental health chatbots

The vast heterogeneity of measurement methods, technology features, and psychotherapeutic approaches across the available research makes it challenging to compare and draw conclusions about the effectiveness of chatbots in the mental health field. However, the most noticeable benefit of mental health chatbots in general and clinical participants is empirically shown in depression over other disorders. This was possibly due to the larger number of studies explicitly developing their chatbots for depressive symptoms and publicly reporting outcomes of them, given the fact that depression is one of the most prevalent mental disorders (Ferrari et al., 2014). In contrast, the results on anxiety and stress, which were the second most targeted issues in the reviewed literature, are still inconclusive as some studies found more remarkable improvement from chatbots (e.g., Fulmer et al., 2018; Liu et al., 2022) while some did not (e.g., Fitzsimmons-Craft et al., 2021; Klos et al., 2021). The unclear effectiveness of chatbots on anxiety and stress, compared to depression, may be the result of a lack of consistency in outcome measures used. Most depression scores in the assessed literature were derived from the PHQ-9 measure, which is highly sensitive to symptom change (Sun et al., 2020), possibly leading to a more significant outcome. In contrast, anxiety and stress levels were assessed with various scales across studies. Although the results implied that chatbots might not be effective enough for emotion management and psychological well-being, no studies reported lower scores after the interventions in these outcomes. Therefore, the effectiveness of the chatbots on mental health concerns which were not statistically significant, but had the direction of a positive change within this review, could yield greater between-group differences if more future research with the standardised ways of measurement is undertaken. Additionally, the mental health chatbot seems to be effective for sleep-related problems, eating disorders, panic disorders, and ADHD (Fitzsimmons-Craft et al., 2021; Jang et al., 2021; Oh et al., 2020; Philip et al., 2022), but with a dearth of studies on these focuses, the effectiveness of chatbot intervention still needs further exploration.

4.4. General discussion

Despite few studies investigating these performance domain associations, the relationships between engagement, user attitude, and effectiveness of mental health chatbots were established in the body of considered literature. Klos et al. (2021) found an impact of positive feedback on increased interactions with the chatbot, suggesting that those who view it as applicable would continue the use. This link could also be supported by the accompanying results of perceived usefulness and reuse intention, which implies future increased engagement, in two studies (Dosovitsky & Bunge, 2023; Greer et al., 2019). Moreover, it could be suggested that sufficient engagement is crucial for chatbot interventions to facilitate significant psychological improvement. This is because several studies found a positive association between these two factors (Greer et al., 2019; Hunt et al., 2021; Jang et al., 2021) and significant mental improvements only among participants adhering to the chatbot enough (Ly et al., 2017). These findings may be partially explained by the combination of the Technology Acceptance Model (Davis, 1989) and the therapeutic alliance model (Ardito & Rabellino, 2011), proposing that user attitude affects the degree of their engagement with chatbots which results in interventions' effectiveness. However, it could also be the case that positive mental outcomes from chatbots motivate users to continue their use and favour the intervention. With unclear directions and scarce evidence of their associations in the present review, a further attempt to comprehensively investigate chatbots should be encouraged to provide a better understanding.

Individual differences may partly influence the performance of mental health chatbots in the three domains of engagement, user attitude, and effectiveness. According to Hunt et al. (2021), Iglesias et al. (2022), and Lavelle et al. (2022), individuals with higher psychological risks and issues were more likely to be interested in and adhere to chatbot intervention. One explanation for these findings may be that people with noticeable problems might be more aware of the specific services they require and perceive chatbots to be useful for meeting these requirements, consequently motivating them to keep interacting with the chatbot. Though chatbot interventions seemed to be less efficient for older people, according to their high dropout and low acceptability (Loveys et al., 2021; Stara et al., 2021), the successful results from Bennoin et al. (2022) using senior samples could imply that it is not always the case. Moreover, Stara et al. (2021) found more negative perceptions in those with more cognitive impairment and less digital experience. Loveys et al. (2021) also reported that the dropout reasons of their participants were usually related to technical difficulties, and the remaining samples could complete the procedure independently after training. Hence, factors that possibly affect the performance of chatbots are not directly age but rather technology familiarity, which is commonly less in older populations.

The embodied psychological method can be another factor driving variances in chatbot usage and impacts. CBT, heavily applied in many chatbot-based interventions, has been supported as an effective technique for several mental illnesses in traditional and computer-mediated treatments (Andrews et al., 2018; David et al., 2018). Nonetheless, the strict therapist-oriented structure of it (Lee et al., 2013), combined with chatbots' lower adaptability than psychologists, could limit the scope of chatbots' responses as this restriction was usually evident in users' unfavourable experiences with chatbots primarily using CBT (e.g., Jang et al., 2021; Liu et al., 2022). Two studies also found that chatbots with other well-constructed approaches, such as MOL and motivation interviews, positively affected user attitude and resulted in higher engagement, reuse motivation, and effectiveness than chatbots with low-constructed techniques (Park et al., 2019; Bennion et al., 2020). However, determining which therapeutic frameworks are suitable for chatbot implementations is challenging because most mental health chatbots are embodied with several rather than a single technique.

The quality of technology in the interventions appears to be improving, as a growing proportion of chatbot responses in this review were derived from advanced language processing of AI operation (e.g., He et al., 2022; Liu et al., 2022), allowing users to have a more profound and natural conversation with it. However, these enhanced capabilities cannot completely alleviate the rule-based chatbot's issues, such as the repetitive and absurd responses, because these problems were still widely reported by users of both chatbot types. Furthermore, AI implementation resulted in novel disadvantages in user perception that did not exist in the rule-based chatbots, namely uncanny valley, due to the overwhelming human-like ability of it (Ta et al., 2020). The virtual agent was another technology increasingly used in the reviewed chatbot interventions. According to the findings of Loveys et al. (2021) and Martnez-Miranda et al. (2019), while the human-like features of their agent were moderately acceptable, the robotic voice used was a key reason for a lack of engagement among users. Similarly, emojis included initially in various chatbot outputs to deliver more meaningful responses were seen negatively by young users due to their ambiguous meaning (Dosovitsky & Bunge, 2023). Overall, the technologies mentioned above should be cautiously integrated into the mental health chatbot because they may either aid or obstruct the psychotherapeutic procedure.

4.4.1. Practical implications

The assessed chatbot seemed to perform better than the compared approaches in terms of engagement, user attitude, and effectiveness, and even comparable to the human-operated therapy in one study (Leo et al., 2022)—moreover, no investigated papers reported later worsening symptoms. Hence, the chatbot may be safe enough to utilise for mental health gain in the general and clinical population with a low degree of severity, especially in the case of depression. Because individuals with higher health issues often showed greater engagement with the chatbot, it could be utilised to encourage people with stigma who have not yet received treatment to seek professional care, as suggested by Miles et al. (2021). Given the deficient understanding of AI's potential and threats, adding personalisation features (e.g., reminders and recommendations) and expanding validated content may be efficient ways to improve chatbot interventions. The promising success of mental health chatbots across countries could imply that they can be used globally. However, implementing them in vulnerable populations, such as those with severe physical/cognitive impairments and low digital competency, may be inappropriate until better digital transformations are prepared since most participants in this review had similar backgrounds in moderate-to-high education and technology familiarity. Combining the findings that chatbots with rigorous approaches reduced symptoms more successfully than those with less psychotherapeutic intensity and users preferred to obtain more relevant responses to their unique problems, it could be suggested that the chatbot should be specifically developed and embodied with unique content for each symptom, rather than providing broad but unfocused counselling.

4.4.2. Research implications

To successfully implement mental health chatbots, it is crucial to achieve positive engagement, user attitude, and effectiveness, which their links were demonstrated in this review. Hence, conducting comprehensive studies that evaluate all three performance areas simultaneously should be encouraged. Additionally, there is a need for more studies on a broader range of mental health issues, specific demographic conditions, such as adolescents or seniors, and larger sample sizes since the lack of such evidence was found in the current review. Another gap in the literature that needs to be addressed is the appropriate way to integrate traditional psychological knowledge, such as CBT and therapeutic alliances, into digital contexts via chatbot technology. Future research should also use a standard set of verified measurements and outcome units to facilitate cross-study comparison and interpretation. Most investigations in the current study were completed within a month. Therefore, researchers should lengthen the study to determine if the positive potential of mental health chatbots seen in the short term can sustain over time. Little research in this review investigated popular chatbots, such as Woebots, Tess (now known as 'Cass') and Wysa. Thus, these widely used chatbots should be properly evaluated for the health and safety of the majority.

4.4.3. Limitations

The inclusion criteria were limited to chatbot performance evaluations from peer-reviewed articles with comparators for effectiveness assessment to ensure a certain level of evidence quality. Consequently, it was likely that findings from grey literature and other sources that could contribute important data were overlooked. The validity and generalizability of the findings to real-world situations may be limited due to the shortage of included studies exploring mainstream chatbots and the exclusion of single-arm effectiveness research, which is typical for gaining real-world information. Owing to the lack of standardised definitions for several terms under the user attitude aspect of digital interventions, other chatbot evaluations employing different theoretical frameworks may reach different conclusions from the present review. For example, usability, categorised as a separate term in this review, is a subcategory under the acceptability umbrella term according to the Technology Acceptance Model (Davis, 1989). Finally, Due to the limited number of chatbots considered in these conclusions and the rapid development of digital technology, regularly conducting systematic reviews in the field is necessary.

5. Conclusion

Evidence from the present review suggests that engagement with mental health chatbots, which generally meet the standard of digital intervention but not traditional ones, can still lead to desired outcomes. Users commonly have positive attitudes towards chatbots regarding user experiences and acceptability but are still uncertain about their usability. Whereas the effectiveness of these interventions across populations and targeted issues has been demonstrated, the conflicting and limited findings on some demographics and symptoms suggest further evaluation. Based on the available evidence, engagement, user attitude, and effectiveness of mental health chatbots appear positively associated with each other. Therefore, more thorough effort in evaluating them, especially the widely used ones, is recommended for effective intervention development. Finally, since the diversity in psychotherapeutic approaches and featured technologies can affect these three performances, pertinent recommendations for real-world implementation and future research have been discussed in this review.

CRediT authorship contribution statement

Sucharat Limpanopparat: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft. **Erin Gibson:** Data curation, Formal analysis, Methodology, Project administration, Writing – review & editing. **Dr Andrew Harris:** Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors have no conflict of interest to declare.

Appendix 1. Characteristics of included studies

In summary, the current review included 31 studies published between 2017 and 2023. The studies took place in a diverse range of countries, with the most in the US (n = 13), and two studies were produced across more than one nation. Almost half of the studies (15/31) recruited participants from nonclinical adults with either an interest in chatbot use or some degree of psychological risks. In contrast, 12 studies used clinical populations with psychological or physical health issues. The most frequently targeted mental health concerns were depression and anxiety (n = 8) and psychological well-being (n = 7), respectively. Psychological content and process in two-thirds of chatbots were constructed based on the CBT framework. The chatbots were presented as a standalone application (n = 16), web-based platform (n = 4), or multimodalities via social media message feature or SMS (n = 11). Nineteen chatbot responses were drawn from a rule-based technique. In contrast, the others (n = 12) utilised AI to perform natural language processing in comprehending and generating responses. The chatbot led the conversation in 23 studies, by the user in 2 and both in 6 studies. Users in most studies interacted with the chatbots only through typing text (n = 24), while users in others could also use emojis (n = 7) or their voice (n = 3). Eighteen chatbots solely conveyed text-based communications, but 13 could interact with users in several ways, including speech (n = 9). Only six chatbots had a virtual representation, with three of them being able to show virtual movements (facial expressions and body language).

Authors (year)	Country	Research design	Sample populations	Chatbot name	Target issues	Psychological technique	Platform	Response	Dialogue intuitive	Input	Output	Visual agent	Research quality
Lavelle et al. (2022)	Ireland	RCT	General adults	N/A	Negative Self-Referential Thoughts	Cognitive Defusion (ACT) or Cognitive Restructuring (CBT)	Multimodal platform (via Facebook message)	Rule-based	Chatbot	Text, emoji	Multi (text, emoji, voice, video)	NO	Med
Zhu et al. (2021).	China	Quantitative	General adults	Xiaolv	Mental health well-being during COVID-19	Psychological assistance	Multimodal platform (via WeChat)	AI-based	Both	Text	Multi (text, image, voice)	NO	High
Fitzpatrick et al. (2017)	US	Mixed methods	General adults	Woebot	Depression and anxiety	CBT	Stan-alone application	Rules-based	Chatbot	Text, emoji	Multi (text, emoji, image, video)	NO	High
He et al. (2022)	China	Mixed methods	General adults	Xiaoe	Depression	CBT	Stan-alone application	AI-based	Chatbot	Text	Multi (text, image, voice)	No	Med
Medeiros et al. (2022)	13 different countries	RCT	General adults	N/A	Daily-life stress	Emotion regulation strategies	Multimodal platform (via Facebook message)	Rule-based	User	Text, emoji	Text	No	Med
Dosovitsky et al. (2020)	US	Quantitative	General population	Tess	Depression	CBT, emotion-focused therapy, solution-focused therapy, motivational interviewing	Multimodal platform (via Facebook message)	AI-based	Chatbot	Text	Text, emoji	No	Low
Philip et al. (2022).	France	Non-RCT	General adults	KANOPEE	Insomnia	CBT	Stand-alone application	Rule-based	Chatbot	Text	Multi (text, image, voice, virtual movement)	Yes	High
Hunt et al. (2021)	US	RCT	Clinical adults	Zemedy	Gastrointestinal (GI) related anxiety	CBT	Stand-alone application	Rule-based	Chatbot	Text	Text	No	High
Ta et al. (2020)	US	Qualitative	General adults	Replika	Mental health well-being	Social support	Stand-alone application	Al-based	Both	Text	Text	No	High
Fitzsimmons-Craft et al. (2021)	US	RCT	General adults	Tessa	Eating disorders	CBT	Multimodal platform (via SMS or Facebook message)	Rule-based	Chatbot	Text, emoji	Multi (text, emoji, image)	No	Med
Iglesias et al. (2022)	US	Quantitative	Clinical adults	Wysa for Return to Work	Mental health well-ing in work-related injuries	CBT, resilience, recovery-based intervention	Stan-alone application	Al-based	Chatbot	Text	Text	No	Low
Liu et al. (2022)	China	Mixed methods	General adults	Xiaonan	Depression	CBT	Stan-alone application	Al-based	Both	Text, voice	Text	No	Low
Ludin et al. (2022)	New Zealand	Mixed methods	General young people	Aroha	Anxiety	CBT, positive psychology	Multimodal platform (via Facebook message)	Rule-based	User	Text	Multi (text, image, video)	Yes	Low
Oh et al. (2020)	Seoul	RCT	Clinical adult	N/A	Panic disorder	CBT	Stan-alone application	Rule-based	Both	Text	Multi (text, image, voice, video)	No	Med
Nicol et al. (2022)	US	RCT	Clinical young people	W-genz within Woebot	Depression and Anxiety	CBT	Stan-alone application	Rules-based	Chatbot	Text	Multi (text, emoji, image, video)	No	Med
Bennion et al. (2020)	UK	RCT	General older people	MYLO	Depression, Anxiety, and Stress	Method of Levels (MOL) therapy	Web-based	Rule-based	Chatbot	Text	Text	No	High
Dosovitsky and Bunge (2022)	US	Qualitative	General young people	Beth Bot	Depression	CBT	Multimodal platform (via Facebook message)	Rule-based	Chatbot	Text, emoji	Text, emoji	No	High
Park et al. (2019)	Seoul	Qualitative	General adults	Bonobot	Stress	Motivational interviewing	Web-based	Rule-based	Chatbot	Text	Text	No	High
Stara et al. (2021)	Italy, Luxembourg, Netherlands	Mixed methods	Clinical older people	Anne	Dementia	CBT	Stan-alone application	Rule-based	Both	Text, voices	Multi (text, image, voice, video)	Yes	High
Greer et al. (2019)	US	Mixed methods	Clinical adults	Vivobot	Depression and Anxiety in Cancer	Stress and Coping theory, the Broaden-and-Build theory	Multimodal platform (via Facebook message)	Rule-based	Chatbot	Text	Multi (text, image, voice, video)	No	Med

(continued on next page)

(*continued*)

Authors (year)	Country	Research design	Sample populations	Chatbot name	Target issues	Psychological technique	Platform	Response	Dialogue intuitive	Input	Output	Visual agent	Research quality
Suganuma et al. (2018)	Japan	Non-RCT	General adults	SABORI	Mental health well-being	CBT	Web-based	AI-based	Chatbot	Text	Text	No	Low
Fulmer et al. (2018)	US	Mixed methods	General adults	Tess	Depression and Anxiety	CBT,emotion-focused therapy, solution-focused therapy, motivational interviewing	Multimodal platform (via Facebook message)	AI-based	Chatbot	Text	Text, emoji	No	Med
Martínez-Miranda et al. (2019)	Mexico	Mixed methods	Clinical adults	Helpath	Suicidality risks	CBT	Stan-alone application	Rule-based	Both	Text	Multi (text, image, voice, virtual movement)	Yes	Med
Jang et al. (2021)	Seoul	RCT	Clinical adults	Todaki	ADHD	CBT	Stand-alone application	Rule-based	Chatbot	Text	Text, emoji	Yes	Med
Loveys et al. (2021)	New Zealand	Mixed methods	Clinical adults	Bella	Mental health well-being during COVID-19	Relationship building strategies, psychoeducation	Web-based	AI-based	Chatbot	Text, voice	Multi (text, image, voice, virtual movement)	Yes	High
Leo et al. (2022)	US	non-RCT	Clinical adults	Wysa for chronic pain	Depression and Anxiety in chronic pain	CBT, mindfulness, and motivational interviewing	Stan-alone application	AI-based	Chatbot	Text, emoji	Text, emoji	No	High
Park et al. (2022)	US	RCT	General adults	N/A	Mental health well-being	The emotional disclosures, psychoeducation	Stan-alone application	Rule-based	Chatbot	Text	Text, emoji	No	Low
Ly et al. (2017)	Sweden	Mixed methods	General adults	Shim	Mental health well-being	Positive psychology, CBT	Stan-alone application	Rule-based	Chatbot	Text	Text	No	High
Shah et al. (2022)	US	Mixed methods	Clinical adults	Alex	Eating disorders	Psychoeducation with motivational interviewing	Multiplatform (via SMS or Facebook Messenger)	Rule-based	Chatbot	Text	Text	No	High
Sinha et al. (2022)	US	Quantitative study	Clinical adults	Wysa for chronic pain	Depression and anxiety in chronic pain	CBT, mindfulness, motivational interviewing	Stan-alone application	AI-based	Chatbot	Text, emoji	Text, emoji	No	High
Klos et al. (2021)	Argentina	RCT	General adults	Tess	Depression and Anxiety	CBT, emotion-focused therapy, solution-focused therapy, motivational interviewing	Multimodal platform (via Facebook message)	AI-based	Chatbot	Text	Text, emoji	No	High

References

Abd-Alrazaq, A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics, 132*, Article 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

Abd-Alrazaq, A., Alajlani, M., Ali, N., Denecke, K., Bewick, B. M., & Househ, M. (2021). Perceptions and opinions of patients about mental health chatbots: Scoping review. *Journal of Medical Internet Research, 23*(1), Article e17828. https://doi.org/10.2196/17828

Abd-Alrazaq, A., Rababeh, A., Alajlani, M., Bewick, B. M., & Househ, M. (2020). Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis. *Journal of Medical Internet Research, 22*(7), Article e16021. https://doi.org/10.2196/16021

Ahmed, A., Aziz, S., Khalifa, M., Shah, U., Hassan, A., Abd-Alrazaq, A., & Househ, M. (2022). Thematic analysis on user reviews for depression and anxiety chatbot apps: Machine learning approach. *JMIR Formative Research, 6*(3), Article e27654. https://doi.org/10.2196/27654

Andrews, G. E., Basu, A., Cuijpers, P., Craske, M. G., McEvoy, P., English, C., & Newby, J. M. (2018). Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: An updated meta-analysis. *Journal of Anxiety Disorders, 55*, 70–78. https://doi.org/10.1016/j.janxdis.2018.01.001

Ardito, R. B., & Rabellino, D. (2011). Therapeutic alliance and outcome of psychotherapy: Historical excursus, measurements, and prospects for research. *Frontiers in Psychology, 2*. https://doi.org/10.3389/fpsyg.2011.00270

Bennion, M., Hardy, G. E., Moore, R., Kellett, S., & Millings, A. (2020). Usability, acceptability, and effectiveness of web-based conversational agents to facilitate problem solving in older adults: Controlled study. *Journal of Medical Internet Research, 22*(5), Article e16794. https://doi.org/10.2196/16794

Borghouts, J., Eikey, E. V., Mark, G., De Leon, C., Schueller, S. M., Schneider, M., Stadnick, N. A., Zheng, K., Mukamel, D. B., & Sorkin, D. H. (2021). Barriers to and facilitators of user engagement with digital mental health interventions: Systematic review. *Journal of Medical Internet Research, 23*(3), Article e24387. https://doi.org/10.2196/24387

Boucher, E. M., Harake, N., Ward, H., Stoeckl, S. E., Vargas, J., Minkel, J., Parks, A. C., & Zilca, R. D. (2021). Artificially intelligent chatbots in digital mental health interventions: A review. *Expert Review of Medical Devices, 18*(sup1), 37–49. https://doi.org/10.1080/17434440.2021.2013200

Boyd, K., Potts, C., Bond, R., Mulvenna, M., Broderick, T., Burns, C., … Booth, F. (2022). Usability testing and trust analysis of a mental health and wellbeing chatbot. In *Proceedings of the 33rd European conference on cognitive ergonomics* (pp. 1–8).

Brooke, J. H. (1996). *SUS: A "quick and dirty" usability scale* (pp. 207–212). CRC Press eBooks. https://doi.org/10.1201/9781498710411-35

Campion, J., Javed, A., Lund, C., Sartorius, N., Saxena, S., Marmot, M., Pagel, J. M., & Udomratn, P. (2022). Public mental health: Required actions to address implementation failure in the context of COVID-19. *The Lancet Psychiatry, 9*(2), 169–182. https://doi.org/10.1016/s2215-0366(21)00199-1

Cella, D., Yount, S., Rothrock, N. E., Gershon, R., Cook, K. H., Reeve, B. B., Ader, D. N., Fries, J. F., Bruce, B., & Rose, M. (2007). The patient-reported outcomes measurement information system (PROMIS). *Medical Care, 45*(5), S3–S11. https://doi.org/10.1097/01.mlr.0000258615.42478.55

Clay, R. A. (2021). Mental health apps are gaining traction. https://www.apa.org/monitor/2021/01/trends-mental-health-apps.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2008). Developing and evaluating complex interventions: The new medical research Council guidance. *BMJ, a1655*. https://doi.org/10.1136/bmj.a1655

David, D., Cristea, I. A., & Hofmann, S. G. (2018). Why cognitive behavioral therapy is the current gold standard of psychotherapy. *Frontiers in Psychiatry, 9*. https://doi.org/10.3389/fpsyt.2018.00004

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Management Information Systems Quarterly, 13*(3), 319. https://doi.org/10.2307/249008

Dixon, L. B., Holoshitz, Y., & Nossel, I. (2016). Treatment engagement of individuals experiencing mental illness: Review and update. *World Psychiatry, 15*(1), 13–20. https://doi.org/10.1002/wps.20306

Donkin, L., Christensen, H., Naismith, S. L., Neal, B., Hickie, I. B., & Hickie, I. B. (2011). A systematic review of the impact of adherence on the effectiveness of e-therapies. *Journal of Medical Internet Research, 13*(3), e52. https://doi.org/10.2196/jmir.1772

Dosovitsky, G., & Bunge, E. (2022). Development of a chatbot for depression: Adolescent perceptions and recommendations. *Child and Adolescent Mental Health, 28*(1), 124–127. https://doi.org/10.1111/camh.12627

Dosovitsky, G., Pineda, B. S., Jacobson, N. C., Chang, C., Escoredo, M., & Bunge, E. L. (2020). Artificial intelligence chatbot for depression: Descriptive study of usage. *JMIR Formative Research, 4*(11), Article e17065. https://doi.org/10.2196/17065

Ellis-Brush, K. (2021). Augmenting coaching practice through digital methods. *International Journal of Evidence Based Coaching and Mentoring*, 187–197. https://doi.org/10.24384/er2p-4857

Ferrari, A. J., Norman, R. E., Freedman, G., Baxter, A. J., Pirkis, J., Harris, M., Page, A., Carnahan, E., Degenhardt, L., Vos, T., & Whiteford, H. (2014). The burden attributable to mental and substance use disorders as risk factors for suicide: Findings from the global burden of disease study 2010. *PLoS One, 9*(4), Article e91936. https://doi.org/10.1371/journal.pone.0091936

Fitzpatrick, K., Darcy, A. M., & Vierhile, M. (2017). 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 Mental Health, 4*(2), Article e19. https://doi.org/10.2196/mental.7785

Fitzsimmons-Craft, E. E., Chan, W., Smith, A. C., Firebaugh, M. L., Fowler, L. A., Topooco, N., DePietro, B., Wilfley, D. E., Taylor, C. B., & Jacobson, N. C. (2021). Effectiveness of a chatbot for eating disorders prevention: A randomized clinical trial. *International Journal of Eating Disorders, 55*(3), 343–353. https://doi.org/10.1002/eat.23662

Fleming, T., Bavin, L. M., Lucassen, M., Stasiak, K., Hopkins, S., & Merry, S. E. (2018). Beyond the trial: Systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. *Journal of Medical Internet Research, 20*(6), e199. https://doi.org/10.2196/jmir.9275

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (TESS) to relieve symptoms of depression and anxiety: Randomized controlled trial. *JMIR Mental Health, 5*(4), Article e64. https://doi.org/10.2196/mental.9782

Gabrielli, S., Rizzi, S., Bassi, G., Carbone, S., Maimone, R., Marchesoni, M., & Forti, S. (2021). *Engagement and effectiveness of a healthy coping intervention via chatbot for university students: Proof-of-concept study during the COVID-19 pandemic (preprint)*. Jmir Mhealth and Uhealth. https://doi.org/10.2196/27965

Gabrielli, S., Rizzi, S., Carbone, S., & Donisi, V. (2020). A chatbot-based coaching intervention for adolescents to promote life skills: Pilot study. *JMIR Human Factors, 7*(1), Article e16762. https://doi.org/10.2196/16762

Gaffney, H., Mansell, W., & Tai, S. (2019). Conversational agents in the treatment of mental health problems: Mixed-method systematic review. *JMIR Mental Health, 6*(10), Article e14166. https://doi.org/10.2196/14166

Greer, S., Ramo, D. E., Chang, Y. J., Fu, M., Moskowitz, J. T., & Haritatos, J. (2019). 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, 7*(10), Article e15018. https://doi.org/10.2196/15018

Griffin, A., Xing, Z., Khairat, S., Wang, Y., Bailey, S. C., Arguello, J., & Chung, A. E. (2020). Conversational agents for chronic disease self-management: A systematic review. *American Medical Informatics Association Annual Symposium, 2020*, 504–513. https://europepmc.org/article/MED/33936424.

He, Y., Yang, L., Zhu, X., Wu, B., Zhang, S., Qian, C., & Tian, T. (2022). Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: Single-blind, three-arm randomized controlled trial. *Journal of Medical Internet Research, 24*(11), Article e40719. https://doi.org/10.2196/40719

Hong, Q., Pluye, P., Bujold, M., & Wassef, M. (2017). Convergent and sequential synthesis designs: Implications for conducting and reporting systematic reviews of qualitative and quantitative evidence. *Systematic Reviews, 6*(1). https://doi.org/10.1186/s13643-017-0454-2

Hunt, M., Miguez, S. M., Dukas, B., Onwude, O., & White, S. (2021). Efficacy of zemedy, a mobile digital therapeutic for the self-management of irritable bowel syndrome: Crossover, randomized controlled trial. *Jmir Mhealth and Uhealth, 9*(5). https://doi.org/10.2196/26152

Iglesias, M., Sinha, C., Vempati, R., Grace, S. E., Roy, M., Chapman, W. C., & Rinaldi, M. L. (2022). Evaluating a Digital Mental health Intervention (WYSA) for workers' compensation claimants. *Journal of Occupational and Environmental Medicine, 65*(2), e93–e99. https://doi.org/10.1097/jom.0000000000002762

IMARC Group. (2022). Healthcare chatbots market share, size and forecast 2023-2028. https://www.imarcgroup.com/healthcare-chatbots-market.

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *Jmir Mhealth and Uhealth, 6*(11), Article e12106. https://doi.org/10.2196/12106

International Organization for Standardization. (2013). *ISO/TS 20282-2:2013 usability of consumer products and products for public use — Part 2: Summative test method*. ISO. https://www.iso.org/standard/62733.html.

Jang, S., Kim, J., Kim, S., Hong, J., Kim, S., & Kim, E. (2021). Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: A development and feasibility/usability study. *International Journal of Medical Informatics, 150*, Article 104440. https://doi.org/10.1016/j.ijmedinf.2021.104440

Kamita, T., Ito, T., Matsumoto, A., Munakata, T., & Inoue, T. (2019). A chatbot system for mental healthcare based on SAT counseling method. *Mobile Information Systems, 2019*, 1–11. https://doi.org/10.1155/2019/9517321

Klos, M. C., Escoredo, M., Joerin, A., Lemos, V., Rauws, M., & Bunge, E. L. (2021). Artificial intelligence–based chatbot for anxiety and depression in university students: Pilot randomized controlled trial. *JMIR Formative Research, 5*(8), Article e20678. https://doi.org/10.2196/20678

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine, 16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Lallemand, C., Gronier, G., & Koenig, V. (2015). User experience: A concept without consensus? Exploring practitioners' perspectives through an international survey. *Computers in Human Behavior, 43*, 35–48. https://doi.org/10.1016/j.chb.2014.10.048

Lau, N., O'Daffer, A., Colt, S., Yi-Frazier, J. P., Palermo, T. M., McCauley, E., & Rosenberg, A. R. (2020). Android and iPhone mobile apps for psychosocial wellness and stress management: Systematic search in app stores and literature review. *Jmir Mhealth and Uhealth, 8*(5), Article e17798. https://doi.org/10.2196/17798

Lavelle, J., Dunne, N., Mulcahy, H. E. M., & McHugh, L. (2022). Chatbot-delivered cognitive defusion versus cognitive restructuring for negative self-referential thoughts: A pilot study. *Psychological Record, 72*, 247–261. https://doi.org/10.1007/s40732-021-00478-7

Lederman, R., & D'Alfonso, S. (2021). The digital therapeutic alliance: Prospects and considerations. *JMIR Mental Health, 8*(7), Article e31385. https://doi.org/10.2196/31385

Lee, E. (2012). Review of the psychometric evidence of the perceived stress scale. *Asian Nursing Research, 6*(4), 121–127. https://doi.org/10.1016/j.anr.2012.08.004

Lee, J. Y. V., Neimeyer, G. J., & Rice, K. (2013). The relationship between therapist epistemology, therapy style, working alliance, and interventions use. *American Journal of Psychotherapy, 67*(4), 323–345. https://doi.org/10.1176/appi.psychotherapy.2013.67.4.323

Leo, A. J., Schuelke, M. J., Hunt, D., Miller, J. P., Areán, P. A., & Cheng, A. L. (2022). Digital mental health intervention plus usual care compared with usual care only and usual care plus in-person psychological counseling for orthopedic patients with symptoms of depression or anxiety: Cohort study. *JMIR Formative Research, 6*(5), Article e36203. https://doi.org/10.2196/36203

Liang, J., Xian, D., Liu, X., Fu, J., Zhang, X., Tang, B., & Lei, J. (2018). Usability study of mainstream wearable fitness devices: Feature analysis and system usability scale evaluation. *Jmir Mhealth and Uhealth, 6*(11), Article e11066. https://doi.org/10.2196/11066

Lim, S. Y., Shiau, C., Cheng, L., & Lau, Y. (2022). Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: A systematic review and meta-regression. *Behavior Therapy, 53*(2), 334–347. https://doi.org/10.1016/j.beth.2021.09.007

Liu, H., Peng, H., Song, X., Xu, C., & Zhang, R. (2022). Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness. *Internet Interventions, 27*, Article 100495. https://doi.org/10.1016/j.invent.2022.100495

Loveys, K., Sagar, M., Pickering, I., & Broadbent, E. (2021). 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 Mental Health, 8*(11), Article e31586. https://doi.org/10.2196/31586

Lovibond, S. H., & Lovibond, P. F. (1995). *Depression anxiety stress scales (DASS–21, DASS–42) [database record]*. APA PsycTests. https://doi.org/10.1037/t01004-000

Ludin, N., Holt-Quick, C., Hopkins, S., Stasiak, K., Hetrick, S., Warren, J., & Cargo, T. (2022). A chatbot to support young people during the COVID-19 pandemic in New Zealand: Evaluation of the Real-World rollout of an open trial. *Journal of Medical Internet Research, 24*(11), Article e38743. https://doi.org/10.2196/38743

Ly, K. H., Ly, A., & Andersson, G. (2017). A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. *Internet Interventions, 10*, 39–46. https://doi.org/10.1016/j.invent.2017.10.002

Manh-Tung, H., Le, N. B., Manh-Toan, H., Mantello, P., & Ghotbi, N. (2023). Understanding the acceptance of emotional artificial intelligence in Japanese healthcare system: A cross-sectional survey of clinic visitors' attitude. *Technology in Society, 72*, Article 102166. https://doi.org/10.1016/j.techsoc.2022.102166

Martínez-Miranda, J., Martínez, A., Ramos, R., Aguilar, H., Jimenez, L. M., Arias, H., Rosales, G., & Valencia, E. (2019). Assessment of users' acceptability of a mobile-based embodied conversational agent for the prevention and detection of suicidal behaviour. *Journal of Medical Systems, 43*(8). https://doi.org/10.1007/s10916-019-1387-1

McTear, M. F., Callejas, Z., & Griol, D. (2016). *The conversational interface: Talking to smart devices*. https://link.springer.com/10.1007/978-3-319-32967-3.

Medeiros, L., Bosse, T., & Gerritsen, C. (2022). Can a chatbot comfort humans? Studying the impact of a supportive chatbot on users' self-PErceived stress. *IEEE Transactions on Human-Machine Systems, 52*(3), 343–353. https://doi.org/10.1109/thms.2021.3113643

Miles, O., West, R., & Nadarzynski, T. (2021). Health chatbots acceptability moderated by perceived stigma and severity: A cross-sectional survey. *Digital Health, 7*, Article 205520762110630. https://doi.org/10.1177/20552076211063012

Mol, M., Van Schaik, D. J. F., Dozeman, E., Ruwaard, J., Vis, C., Ebert, D. S., Etzelmueller, A., Mathiasen, K., Moles, B., Mora, T., Pedersen, C. P., Skjøth, M. M., Pensado, L. P., Piera-Jimenez, J., Gokcay, D., Ince, B.Ü., Russi, A., Sacco, Y., Zanalda, E., … Smit, J. (2020). Dimensionality of the system usability scale among professionals using internet-based interventions for depression: A confirmatory factor analysis. *BMC Psychiatry, 20*(1). https://doi.org/10.1186/s12888-020-02627-8

Ng, M., Firth, J., Minen, M. T., & Torous, J. (2019). User engagement in mental health apps: A review of measurement, reporting, and validity. *Psychiatric Services, 70*(7), 538–544. https://doi.org/10.1176/appi.ps.201800519

Nicol, G., Wang, R., Graham, S., Dodd, S., & Garbutt, J. (2022). Chatbot-delivered cognitive behavioral therapy in adolescents with depression and anxiety during the COVID-19 pandemic: Feasibility and acceptability study. *JMIR Formative Research, 6*(11), Article e40242. https://doi.org/10.2196/40242

Nwosu, A. N., Boardman, S., Husain, M. M., & Doraiswamy, P. M. (2022). Digital therapeutics for mental health: Is attrition the Achilles heel? *Frontiers in Psychiatry, 13*. https://doi.org/10.3389/fpsyt.2022.900615

Oh, J., Jang, S., Kim, H., & Kim, J. (2020). Efficacy of mobile app-based interactive cognitive behavioral therapy using a chatbot for panic disorder. *International Journal of Medical Informatics, 140*, Article 104171. https://doi.org/10.1016/j.ijmedinf.2020.104171

Oladeji, B. D., & Gureje, O. (2016). Brain drain: A challenge to global mental health. *BJPsych International, 13*(3), 61–63. https://doi.org/10.1192/s2056474000001240

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T., Mulrow, C. D., Shamseer, L., Tetzlaff, J., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, n71*. https://doi.org/10.1136/bmj.n71

Park, S., Choi, J., Lee, S., Oh, C., Kim, C., La, S., Lee, J., & Suh, B. (2019). Designing a chatbot for a brief motivational interview on stress management: Qualitative case study. *Journal of Medical Internet Research, 21*(4), Article e12231. https://doi.org/10.2196/12231

Park, G., Chung, J., & Lee, S. (2022). Effect of AI chatbot emotional disclosure on user satisfaction and reuse intention for mental health counseling: A serial mediation model. *Current Psychology*. https://doi.org/10.1007/s12144-022-03932-z

Perski, O., Crane, D., Beard, E., & Brown, J. (2019). Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. *Digital Health, 5*, Article 205520761988067. https://doi.org/10.1177/2055207619880676

Philip, P., Dupuy, L., Sagaspe, P., De Sevin, E., Auriacombe, M., Taillard, J., Micoulaud-Franchi, J., & Morin, C. M. (2022). Efficacy of a smartphone-based virtual companion to treat insomniac complaints in the general population: Sleep diary monitoring versus an internet autonomous intervention. *Journal of Clinical Medicine, 11*(15), 4387. https://doi.org/10.3390/jcm11154387

Rapp, A., Curti, L., & Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies, 151*, Article 102630. https://doi.org/10.1016/j.ijhcs.2021.102630

Roach, A., Cullinan, S., Shafran, R., Heyman, I., & Bennett, S. (2023). Implementing brief and low-intensity psychological interventions for children and young people with internalizing disorders: A rapid realist review. *British Medical Bulletin, 145*(1), 120–131. https://doi.org/10.1093/bmb/ldad001

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology, 57*, 493–502.

Sekhon, M., Cartwright, M., & Francis, J. J. (2017). Acceptability of healthcare interventions: An overview of reviews and development of a theoretical framework. *BMC Health Services Research, 17*(1). https://doi.org/10.1186/s12913-017-2031-8

Shah, J., DePietro, B., D'Adamo, L., Firebaugh, M. L., Laing, O., Fowler, L. A., Smolar, L., Sadeh-Sharvit, S., Taylor, C. B., Wilfley, D. E., & Fitzsimmons-Craft, E. E. (2022). Development and usability testing of a chatbot to promote mental health services use among individuals with eating disorders following screening. *International Journal of Eating Disorders, 55*(9), 1229–1244. https://doi.org/10.1002/eat.23798

Sinha, C., Cheng, A. L., & Kadaba, M. (2022). Adherence and engagement with a cognitive behavioral therapy–based conversational agent (Wysa for chronic pain) among adults with chronic pain: Survival analysis. *MIR Formative Research, 6*(5), Article e37302. https://doi.org/10.2196/37302

Smith, A. C., Thomas, E., Snoswell, C. L., Haydon, H. M., Mehrotra, A., Clemensen, J., & Caffery, L. J. (2020). Telehealth for global emergencies: Implications for coronavirus disease 2019 (COVID-19). *Journal of Telemedicine and Telecare, 26*(5), 309–313. https://doi.org/10.1177/1357633x20916567

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder. *Archives of Internal Medicine, 166*(10), 1092. https://doi.org/10.1001/archinte.166.10.1092

Stara, V., Vera, B., Bolliger, D., Paolini, S., De Jong, M., Felici, E., Koenderink, S., Rossi, L., Von Doellen, V., & Di Rosa, M. (2021). Toward the integration of technology-based interventions in the care pathway for people with dementia: A cross-national study. *International Journal of Environmental Research and Public Health, 18*(19), Article 10405. https://doi.org/10.3390/ijerph181910405

Suganuma, S., Sakamoto, D., & Shimoyama, H. (2018). An embodied conversational agent for unguided internet-based cognitive behavior therapy in preventative mental health: Feasibility and acceptability pilot trial. *JMIR Mental Health, 5*(3), Article e10454. https://doi.org/10.2196/10454

Sun, Y., Fu, Z., Bo, Q., Mao, Z., Ma, X., & Wang, C. (2020). The reliability and validity of PHQ-9 in patients with major depressive disorder in psychiatric hospital. *BMC Psychiatry, 20*(1). https://doi.org/10.1186/s12888-020-02885-6

Ta, V. P., Griffith, C., Boatfield, C. R., Wang, X., Civitello, M., Bader, H. J., DeCero, E., & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *Journal of Medical Internet Research, 22*(3), Article e16235. https://doi.org/10.2196/16235

Tremain, H., McEnery, C., Fletcher, K., & Murray, G. (2020). The therapeutic alliance in digital mental health interventions for serious mental illnesses: Narrative review. *JMIR Mental Health, 7*(8), Article e17204. https://doi.org/10.2196/17204

Vaidyam, A., Wisniewski, H., Halamka, J., Kashavan, M. S., & Torous, J. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *Canadian Journal of Psychiatry, 64*(7), 456–464. https://doi.org/10.1177/0706743719828977

Wasil, A. R., Palermo, E. H., Lorenzo-Luaces, L., & DeRubeis, R. J. (2021). Is there an app for that? A review of popular apps for depression, anxiety, and well-being. *Cognitive and Behavioral Practice, 29*(4), 883–901. https://doi.org/10.1016/j.cbpra.2021.07.001

Watson, D. I., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Wu, A., Scult, M. A., Barnes, E. R., Betancourt, J. A., Falk, A., & Gunning, F. M. (2021). Smartphone apps for depression and anxiety: A systematic review and meta-analysis of techniques to increase engagement. *Npj Digital Medicine, 4*(1). https://doi.org/10.1038/s41746-021-00386-8

Zhu, Y., Janssen, M., Wang, R., & Liu, Y. (2021). It is me, chatbot: Working to address the COVID-19 outbreak-related mental health issues in China. User experience, satisfaction, and influencing factors. *International Journal of Human-computer Interaction, 38*(12), 1182–1194. https://doi.org/10.1080/10447318.2021.1988236