Current Psychology (2025) 44:13619–13640
https://doi.org/10.1007/s12144-025-08094-2

Chatbots and mental health: a scoping review of reviews

Eric Mayor[1]

Accepted: 8 June 2025 / Published online: 28 June 2025

Abstract

The use of chatbots for mental health is now frequently discussed as a means to increase access to mental health resources. For this scoping review of reviews on chatbots for mental health, we performed a systematic search of the literature on Scopus, Web of Science, PubMed and Dimensions.ai and identified 14 relevant reviews published in scientific journals which identified publications following a systematic search of two or more databases. Three additional relevant reviews were included following forward and backward reference examination. The 17 included reviews are: eight systematic reviews (three with meta-analysis), seven scoping reviews and two unlabeled reviews. We have summarized the scope of the reviews as well as their findings. Most reviews included at least one study with 15 participants or fewer. Few reviews report a significant proportion of randomized controlled trials and some original publications most of the time report on piloting studies. Overall, the reviews examine the opinions of users of chatbots, the features of chatbots, the outcomes and measures in studies relying on chatbots, the effectiveness of chatbots (particularly the meta-analyses) in alleviating mental disorder symptomatology as well as their potential for the assessment of mental health. The use of machine learning and natural language processing frameworks (Dialogflow, RASA) seems to drive an increasing number of studies of chatbots for mental health, but progress is necessary in several areas affecting interaction dynamics.

Keywords Chatbots · Conversational agents · Mental health · Review · Scoping

There has been an important upsurge in prevalence and incidence of mental disorders in the last decades: Globally, the disability-adjusted life years lost to mental disorders has increased from 80.8 million in 1990 to 125.3 million in 2021 (55% increase) most importantly due to depressive disorders (GBD 2019 Mental Disorders Collaborators, 2022). A majority of individuals presenting with mental disorders do not receive mental health services (Henderson et al. 2013; WHO, 2021), and data from 21 countries show that only 9.8% of individuals with anxiety disorders may receive adequate treatment, while 72.4% receive no treatment at all (Alonso et al., 2018). The reasons for such low coverage include on the one hand the lack of availability of mental health professionals: globally there are 3.4 mental health workers for 100,000 people, and 0.01 in low-income countries, 0.4 in lower-middle income countries (WHO, 2021). Other reasons include: stigma and discrimination of help seeking individuals for mental health issues, distance and safety of travel, particularly for members of underrepresented social groups, language barriers, lack of insurance coverage, lack of trust in the healthcare system, lack of awareness about mental health services, critical perceptions about mental health care, and lack of perceived need for care (e.g., Alonso et al., 2018; Cyr et al., 2019; Thornicroft, 2008; Williams et al., 2021).

Technology-assisted and technology-delivered psychotherapy have been considered as potential solutions to this issue. In the recent years, the importance of web-based psychotherapy has risen in both the academic and private sectors (e.g., Esfandiari et al., 2021; Karyotaki et al., 2021; Zhang et al., 2023). Meta-analyses show that the effectiveness of such tools (e.g., iCBT) is sometimes even comparable with face-to-face psychotherapy (Esfandiari et al., 2021; Karyotaki et al., 2021; Zhang et al., 2023). The use of chatbots for psychotherapy has been a research focus since the early test of the Eliza chatbot in 1966 (Weizenbaum, 1966), which was designed to emulate Rogerian

✉ Eric Mayor
ericmarcel.mayor@unibas.ch

[1] Division of Clinical Psychology and Epidemiology, University of Basel, Missionsstrasse 62A, 4055 Basel, Switzerland

psychotherapy as a means to study language interactions rather than psychotherapy. Since then, the technological capabilities have drastically increased, leading to chatbots achieving bonds with clients (e.g., working alliance) comparable to those established with clinicians (Darcy et al., 2021). Interventions, such as cognitive behavioural therapy (CBT), partially or completely delivered through chatbots are considered an important means to reduce access issues (Moore & Caudill, 2019).

In the recent years, a number of reviews of the literature have been written on chatbots for mental health. The aim of this present scoping review is hence to summarize the scope and findings of existing reviews on chatbots for mental health that rely upon systematic searches of the literature (e.g., scoping reviews, systematic reviews and meta-analyses) and are published in scientific journals.

Chatbots for mental health

Chatbots, also known as conversational agents, are applications designed to emulate human interaction with users through natural language (Hussain et al., 2019). The prospect of computer dialogue systems revolutionizing psychiatry is not new (e.g., see Zarr, 1984). However, aside from the early popular chatbot Eliza (Weizenbaum, 1966), until the two last decades, chatbots emulating psychotherapists have not seen as much success as they have in other areas. This is not to say there weren't attempts before then. Slack and Slack (1977) tested a computer dialogue system with patients with mental disorders and reported that patients who talked with the computer were better able to disclose information to a psychotherapist afterward. Bloom et al. (1978) created the Converse language to design early clinical chatbots. A review by Slack (2000) mentions several studies with computerized clinical interviews as well as examples of early chatbots created in Slack's lab and the study by Selmi et al. (1990) which involved depressed patients' conversations with a computer program which was developed using the Converse language and ran CBT sessions through teaching on the basis of the patients' input (rather than really interactive dialogue). Osgood-Hynes et al. (1998) tested an interactive voice response system which provided psychotherapy to depressed patients (in association with learning materials), achieving a 50% reduction in depressive symptoms. More recently, Cole et al. (2003) reported on a chatbot targeted at children with autism spectrum disorder among other chatbots mentioned in the publication. Grolleman et al. (2006) developed a chatbot targeting substance use disorder (smoking cessation), Ku et al. (2007)'s chatbot targeted communication skills of patients with schizophrenia and Konstantinidis et al. (2009)'s chatbot aimed at improving the conversational skills of children with autism spectrum disorders.

The present contribution focuses on developments from that period onward. It synthesizes reviews that are published in scientific journals that focus on empirical studies on chatbots for mental health and seeks to answer the following research questions:

- Is the field of chatbots for mental health growing?
- What are the research foci of the reviews in the field?
- Which are the countries from which most studies in the field originate?
- What are the sample sizes of such studies reported in the reviews?
- What are frequent characteristics of the interventions (chatbots)?
- What is the degree of scientific rigor (e.g., use of randomized controlled trials - RCTs) in the field?
- What are the main clinical targets of the studies and the most frequent measures?
- Are the interventions efficient at reducing mental disorder symptomatology and other targets?
- What are users' perceptions about chatbots for mental health?
- Are there evolutions of the field that can be anticipated?

Methods

On October 19 and 20, 2023, we searched on Scopus, Web of Science, PubMed, and Dimensions.ai, using the queries outlined in Table 1.

Inclusion criteria

We included reviews written in English focusing on empirical studies of chatbots in relation to mental health that relied on systematic searches (i.e., using keywords to systematically identify matching records) of at least two academic databases, focusing on any population and without requirements on the comparator groups examined, if any and including studies with any design (e.g., RCT, quasi-experiment) and focus (e.g., symptom reduction, monitoring/detection, features of chatbots, users' perception).

Exclusion criteria

We excluded reviews which were not published in academic journals. We also excluded reviews for which the included studies were not clearly reported (absent from the reference list without mention in supplementary materials), and

Table 1 Search queries

Database	Number of matches	Query
Scopus	75	(TITLE-ABS-KEY ("chatterbot*" OR "chatbot*" OR "conversational agent*") AND TITLE-ABS-KEY ("psychotherapy" OR "therapy" OR "mental health" OR "symptom*" OR "psychiatr*")) AND (LIMIT-TO (LANGUAGE, "English")) AND (LIMIT-TO (DOCTYPE, "re"))
Web of Science	70	"chatterbot*" OR "chatbot*" OR "conversational agent*" (Topic) AND ("psychotherapy" OR "therapy" OR "mental health" OR "symptom*" OR "psychiatr*" (Topic)), refining on Review article;
Dimension.ai	167	("chatterbot*" OR "chatbot*" OR "conversational agent*") AND ("psychotherapy" OR "therapy" OR "mental health" OR "symptom*" OR "psychiatr*") AND ("review");
Pubmed (using the pubmedR package)	63	("chatterbot*" OR "chatbot*" OR "conversational agent*"[Title/Abstract]) AND ("psychotherapy" OR "therapy" OR "mental health" OR "symptom*" OR "psychiatr*"[Title/Abstract]) AND english[LA] AND Review[PT]

publications which included studies which mostly didn't feature chatbots, despite the title and abstract of the reviews.

Screening of publications

Title and abstract screening was assisted by machine learning using the ASReview Lab software (version 1.2) (ASReview Lab Developers, 2022). Eight relevant records and 40 irrelevant records were coded as input (title and abstract). The process was conducted using default parameters (Feature extraction technique: TF-IDF, Classifier: Naïve Bayes, Query strategy: Maximum; Balance strategy: Dynamic resampling – Double). The records were presented for manual labelling based on title and abstract in the order of relevance determined by the algorithm. All articles were manually assessed for relevance within the software, followed by verification in Excel. Articles considered relevant were downloaded, following which eligibility was assessed from the full text. References of included articles at that stage were examined for potential reviews matching the eligibility criteria. Articles citing the included articles were selected for full text examination based on title following a search for 'intitle: review' among citing articles on Google Scholar).

Data extraction

When available, the following information was extracted from the included reviews: the date of publication, the focus of the review, the number of databases searched, the number of publications included in the review, the references and publication year of the included studies, the countries from which the included studies originated, information regarding the design of the included studies (e.g., RCT), the main characteristics of the samples (e.g., sample sizes, age, gender, number of clinical samples), key chatbots features (e.g., standalone/web-based), the goals (e.g., counselling) and/or roles (e.g., therapist) of the studied chatbots, the clinical targets of the studies, the clinical constructs and the related measures, as well as import findings of the reviews and interesting considerations of the authors of the reviews. The most structured information is reported in Tables 2, 3 and 4, and the rest in the form of summaries.

Results

Figure 1 presents the PRISMA 2020 flowchart. The total number of matching publications was 375 (Scopus: 75, Web of Science: 70, Pubmed: 63, Dimensions.ai: 167). After removing duplicates, 215 unique records remained. Publications that were not published in peer-reviewed journals were then removed, leading to 136 articles for initial title and abstract screening. Following this step, 21 records were deemed eligible for full-text screening and were retrieved. Of these, 14 articles were included in the final review; the rest were excluded based on the criteria outlined earlier.

A search for "review" within the reference lists of these 14 articles did not yield additional eligible publications. However, searching for "review" among citing articles via Google Scholar (conducted on November 16, 2023) resulted in the identification of three additional reviews, which were included.

This review reports on a total of 17 reviews, including eight systematic reviews, three of which incorporated a meta-analysis, as well as seven scoping reviews and two unlabeled reviews. All included reviews employed a systematic search strategy and queried at least two academic databases.

These reviews themselves reported on 238 unique publications. Of these, 172 were included in only one review, 29 in two reviews, 21 in three reviews, 7 in four reviews, and 3 in five reviews. Additionally, two publications appeared in seven, nine, and ten reviews, respectively. On average, each review included 22.82 articles *(SD = 16.75; range: 6 to 54)*. Figure 2 displays the number of included publications per year and indicates that few studies were published before 2010. The reviews included 134 studies published between 2018 and 2023, and 104 studies published before 2018.

Table 2 provides information about the search strategy, included studies and sample characteristics. Table 3

Table 2 Search strategy, included studies and information on included studies

	Review focus	Year search / last update	Number of databases queried	Number of publications included	Number of unique studies	Range of publication years of studies included	Number of RCTs	Number with clinical samples	Studies countries	Sample sizes	Age	Gender
Abd-Alrazaq et al. (2021):	opinions of users	2019	7	37	37	2006—2019	1	21	USA (17 studies), Australia, France, the Netherlands (8 studies each) and 11 other countries	N≤50 in 65% of studies, 51–100 in 14%, 101–200 in 16%, > 200 in 5%	mean: 33.4; SD: 15.2; range: 13 to 79	balanced on average
Abd-Alrazaq et al. (2019):	features of chatbots	2019	5	53	53	2007—2019	14	31	USA (23 studies), Japan (6 studies), Australia (4 studies) and 13 other countries	mean: 75.2; range: 4—454	mean: 37.1	55% male overall
Ahmed et al. (2023):	features of chatbots (depression, anxiety)	2022	5	42	42	2015–2022	N/A	N/A	USA (11 studies), India (8 studies), China (4 studies) and 12 other countries	N/A	N/A	N/A
Bendig et al. (2022):	scope and evidence (general)	2022	7	6	6	2018–2022	4	1	USA (2 studies), and one from each of the following: UK, Sweden, Switzerland, Japan	mean: 156.8; SD: 174.52; range: 28—454	28.5; SD; 5.5	Percent male: 90.3%; 9.7%—53.6%
Jabir et al. (2023):	outcomes targeted and measures	2021	5	31	32	2010–2021	13	9	Half of the studies originated from the USA, 22% from the UK, 9% from Japan. The other studies originated from 5 additional countries	N/A	N/A	N/A
Provoost et al. (2017):	embodied conversation agents	2015	5	54	49	2003–2015	N/A	34	N/A	mean: 90.47, SD = 176.17 (computed from Table 1 data in their article); range: 1—1317	ASD % adult samples: 24%; other targets: 100%	N/A
Martin and Richmond (2023):	chatbots for children and their caregivers	2023	3	15	15	2018—2022	N/A	N/A	USA (5 studies), Australia (2 studies), South Korea (2 studies), and the others studies from 7 other countries	mean: 52; range: 5—400	adolescents or children in 9 studies, of caregivers in 3 studies and of both children and caregivers in 2 studies	N/A
Martínez-Miranda (2017):	embodied conversational agents for suicide risk detection and prevention	2017	5	6	9	2015, 2016	3	4	N/A	mean = 3770.8, sd = 10277.5; range: 33 to 31,144	N/A	N/A

Table 2 (continued)

	Review focus	Year search / last update	Number of databases queried	Number of publications included	Number of unique studies	Range of publication years of studies included	Number of RCTs	Number with clinical samples	Studies countries	Sample sizes	Age	Gender
Vaidyam et al. (2019):	psychiatric chatbots	2018	6	10	10	2010—2017	N/A	4	N/A	mean: 62.54; SD: 58.16; range: 4—179	mean: 33.6	Balanced on average, percent male ranged from 0 to 93%
Vaidyam et al. (2020):	psychiatric chatbots (update)	2018	4	7	7	2018—2020	1	N/A	N/A	median: 75; 6—454	mean: 34.29	average: 46% male
Gaffney et al. (2019):	effectiveness of chatbots	2019	5	13	13	2013—2018	8	8	USA (6 studies), UK (5 studies), one study from Sweden and one from Japan	mean: 97.4; 14 – 454	range: 16—75	Average male: 30%
Ogilvie et al. (2022):	chatbots for substance use disorder	2022	10+(use of institutional tool which queries many scientific and publisher's databases)	6	6	2016—2021	1	3 (+1, unclear)	N/A	mean: 77.6; SD: 67.2; range: 17—180	range: 18 to 76	N/A
Otero-González et al. (2024):	chatbots for screening of depression	2023	6	36	36	2013—2023	4	14	USA (9 studies), Australia (4 studies), Spain (4 studies); France, Germany, India (3 studies each), and the other studies from 7 other countries	mean: 130.61; SD = 171.72; range: 7—887	Not systematically reported	4 studies only women, other studies women and men (other information unavailable)
Pacheco-Lorenzo et al. (2021):	chatbots for early detection of psychiatric disorders and neurocognitive disorders	2020	7	17	17	2014—2020	N/A	N/A	N/A	In Studies using self-collected data (12 studies): mean: 89.2; SD = 95.6	8 studies with eldely people, other with any target users (other information unavailable)	N/A
Abd-Alrazaq et al. (2020):	efficacy of chatbots for symptom alleviation	2019	7	12	12	2015—2018	6	6	USA (4 studies), the others from different countries, such as the UK, Sweden, Australia, Turkey, Japan, China among others	mean: 97.75; SD: 121.79; range: 10—454	mean: 31.3	average male: 35%; range: 9.5% to 56%

Table 2 (continued)

	Review focus	Year search / last update	Number of databases queried	Number of publications included	Number of unique studies	Range of publication years of studies included	Number of RCTs	Number with clinical samples	Studies countries	Sample sizes	Age	Gender
Lim et al. (2022)	chatbots for depression and anxiety	2020	7	11	11	2009—2017	11	7	Germany (5 publications), the USA (3), Switzerland (2) and the UK	mean = 129.36, SD = 114.78, range: 14 to 396	all adults, no further information on age	no information on gender
He et al. (2023)	efficacy of chatbots for symptom alleviation	2022	6	32	32	2012—2022	32	10	USA (28%) and the UK (19%). Three studies were from Sweden. The others studies originated from 10 other countries	mean: 121.09; SD: 171.51; range: 14 -. 927	mean: 36.32; SD: 13.44; range: 21.40 to 71.47	average male: 24.59; range: 0—76%

provides information regarding the targeted disorders as well as the main measures of clinical outcomes and other comments on measures. The summaries of the included reviews below provide information on other important aspects of the reviews. They are followed by a synthesis.

Scoping reviews

The scoping review by Abd-Alrazaq et al. (2021) is interested in the opinions of users on chatbots. In the studies reviewed, 35% of chatbots were web-based, while the remainder were standalone applications. The authors mention 86% were rule-based and the remaining used artificial intelligence (AI). In most cases (86%), dialogue initiation was controlled by the chatbot. The authors used thematic analysis to categorize and characterize the studies. They distinguished between different themes which they constructed on a bottom-up fashion starting with codes: usefulness (typically rated as high), ease of use (typically rated high as well), responsiveness (often mixed or neutral perceptions, repetitiveness of responses highlighted), understandability (understanding by chatbots considered high in several studies, while in several others chatbots considered unable to understand the user), acceptability (typically rated as high), attractiveness (generally low to neutral), trustworthiness (generally rated high), enjoyability (generally high), content provided by the chatbot (users typically satisfied but sometimes content is considered superficial), and comparisons of different chatbots or comparisons of chatbots with human professionals (one study showed preference for chatbot vs human, one study found the opposite and in other studies preferences were mixed; participants might prefer embodied and empathetic chatbots but disclose more to non-embodied chatbots). Overall, users valued features such as embodiment, relationship building, real-time feed-back, the possibility to hold conversations daily with a weekly summary and access to a helpline. The added value of speech was not clear to users in all studies but friendly voices seemed important in this respect. Users sometimes felt the chatbots lacked responsiveness and failed to provide detailed or precise feedback. Some users were critical of the ability of the chatbots to understand their verbal and nonverbal input. The authors note that chatbots offer distinct advantages, such as real-time feedback and daily availability, which therapists can not.

The scoping review by Abd-Alrazaq et al. (2019) is interested in the features of mental health chatbots. In 92.5% of the studies, the responses were rule-based, whereas they were relying upon AI in a small minority of the included studies. Most chatbots (70%) operated as standalone software. The authors do not report on clinical instruments and do not describe specific findings related to outcome

Table 3 Main therapeutic frameworks, clinical targets and instruments

	Therapeutic framework mentioned in two studies or more	Main clinical targets	Main mental health instruments, other comments on measures
Abd-Alrazaq et al. (2021):	Not described	Depression (41 studies), anxiety (6 studies), autism (6 studies), any mental disorder (6 studies). Other disorders and outcomes were also targeted	Not described
Abd-Alrazaq et al. (2019):	CBT	The most commonly addresses mental issues were depression (16 studies), autism (10), anxiety disorders including phobias (9), and post-traumatic stress disorder (7)	Outcomes measured included acceptability (36 studies), effectiveness (33 studies), usability (20 studies) and adoption (7 studies)
Ahmed et al. (2023):	CBT	Both depression and anxiety were targeted by 60% of the chatbots, 38% targeted only depression and one study targeted only anxiety	One version of the PHQ (in order of frequency of use: PHQ-9, PHQ-2, PHQ-8) was used in 36% of the studies, the GAD in 12%. The PANAS and other scales were also used (frequency not reported in the review). The authors note 24% of the studies reported on user perceptions and engagement, but they do not report on these findings
Bendig et al. (2022):	CBT	The targets of the interventions included depression (2 studies), anxiety (2 studies), stress (2 studies), wellbeing (2 studies) among others	The studies relied on different instruments, with the PSS (with 4 or 10 items) being the only one used in more than 1 study. Other instruments were the DASS-21, the PHQ-9, the GAD-7, the PANAS, the SWLS, the FS, the BFI (personality), the WHO-5-J, BADS and K10
Jabir et al. (2023):	CBT (from Appendix 4)	The targets were mental wellbeing in 53% of the studies, depression and anxiety in 13%, depression only in 9% and 25% had other targets	The instruments used in more than two studies were: the PHQ (8 or 9 items) – 8 studies; a version of the PANAS—5 studies; GAD-7, a version of the PSS, the STAI, – each 4 studies; problem-related distress and problem resolution, the Working Alliance Inventory – each 3 studies; Jackson Flow state, Rosenberg self-esteem scale, System usability scale each 2 studies. Other instruments assessing mental health or user satisfaction / chatbot usability were also used in one study
Provoost et al. (2017):	CBT	Autism spectrum disorders (26 studies), Depression (10 studies), Anxiety (5 studies), post-traumatic stress disorder (4 studies), psychosis (4 studies); substance use (4 studies)	The most frequent outcomes were: usability (26 studies), behavioral (25 studies), satisfaction (20 studies), and usage (15 studies)
Martin and Richmond (2023):	Not described	The majority of the studies focusing on specific mental disorders targeted autism (4 studies) while ADHD, depression and anxiety, eating disorders and social anxiety were targeted by one study each. Mental health in general was targeted in the other studies. (total = 15)	The outcomes of the studies focused on the usability and acceptability of the chatbots and other outcomes such as enjoyment, accessibility or technical difficulties
Martínez-Miranda (2017):	CBT	Review focuses on suicide prevention, but chatbots have other main targets	Not systematically described. PHQ-9 and BDI-2 mentioned
Vaidyam et al. (2019):	CBT	The (clinical) targets included: wellbeing, PTSD, anxiety and depression, stress, psychoeducation, medication adherence	The measures used in more than one study were: the PHQ-9 (4 studies), the Post-Traumatic Stress Disorder Checklist (2 studies), the Post-Deployment health Assessment (2 studies), a version of the perceived stress scale (PSS-10 in one study, PSS-4 in one study), the Positive and Negative Affect Scale (2 studies). Other measures included the GAD-7, the Flourishing scale, the SWLS, the Short Form Health Survey-12 among others. Several measures employed were self-developed by the original study authors. Engagement metrics were included in most of the studies (e.g., number of days used, number of logins, usage duration) and satisfaction. Several studies included acceptability and/or usability metrics
Vaidyam et al. (2020):	CBT (from publication titles)	The main clinical targets were mostly depression (4 studies) and anxiety (three studies). Other targets included wellbeing (2 studies), broad spectrum of mental disorders (assessment)	The measures used included (non-exhaustive list): the PHQ-9, the HADS, the Hamilton depression rating scale, the GAD-9, the PANAS, Plutchik Suicide Risk Scale, the WHO-5, as well as several measures of users' perceptions. No clinical measures except the PHQ-9 were used in more than one study

Table 3 (continued)

	Therapeutic framework mentioned in two studies or more	Main clinical targets	Main mental health instruments, other comments on measures
Gaffney et al. (2019):	CBT, MOL	Six studies targeted depression (2 of them also anxiety), 3 studies wellbeing (one of the in relation to stress specifically), two studies psychological distress, one study loneliness and one study acrophobia	The instruments measuring clinical outcomes used in more than one study were: the PHQ-9 (4 studies), the HADS (2 studies), the DASS-21 (2 studies), the PANAS (two studies). In addition, user experience was assessed in 11 of the 13 studies
Ogilvie et al. (2022):	CBT	One of the interventions targeted alcohol abuse (through alcohol education and risk assessment), one targeted drugs and opioid abuse (through question answering), two studies targeted problematic drug and alcohol use (delivering CBT inspired intervention), another investigated market leading assistants' responses to addiction-related support requests (questions to the assistant), one investigated participants' opinions about a chatbot targeting alcohol and drug abuse (role of the chatbot is unclear)	Three studies included clinical instruments. The AUDIT-C was included in all of them. The DAST-10, GAD-7, BSC, PHQ-8, CSQ-8 were included in two studies by the same first author
Otero-González et al. (2024):	Not described	Depression	Instruments used in more than one study were: A version of the PHQ (PHQ-9:16 studies, PHQ-4: one study, PHQ-2: one study): 18 studies; the BDI-2: 4 studies, the GAD-7: 6 studies, the BDI-II: 4 studies, the EORTC QOLINFO25: 2 studies, the IDS-C (clinician rated depression scale): 2 studies the SIGH-D: 2 studies (clinician rated depression interview). Other instruments included (non-exhaustive list): the CES-D, the MBI, the SF-36, the PANAS
Pacheco-Lorenzo et al. (2021):	Not described	Seven publications targeted neurocognitive disorders (e.g., functional memory disorder, Alzheimer's disease, dementia or cognitive impairment in general), 7 depression, and 3 other disorders	The instruments used in more than one study were the PHQ-8 or PHQ-9 (4 studies), 3 studies used clinical interviews and the MMSE (2 studies). The other instruments included (non-exhaustive list): the BDI-II, the GAD-7, the Revised Hasegawa Dementia Scales. Two studies used diagnostic interviews rather than questionnaires for validation. The authors mention only one study assessed usability and another acceptability of the use of chatbots in the included publications
Abd-Alrazaq et al. (2020):	CBT (from publication titles)	The clinical targets were: Depression (7 studies), Anxiety (4 studies), Any mental disorder (3 studies), Acrophobia (1 study)	The instruments used in more than one study were: PHQ-9 (4 studies); GAD-7 (2 studies), PSS–10 (2 studies), Kessler K-10 (2 studies), PANAS (2 studies)
Lim et al. (2022)	CBT, PST	Depression, anxiety, other disorders (authors focus meta-analysis on depression)	The BDI (I or II) was used in 7 articles, the PHQ-9 in two articles and other measurements in the remaining
He et al. (2023):	CBT (53% of studies), PST, MOL, VRET	The interventions targeted: generalized anxiety symptoms (34%), depressive symptoms (28%), specific anxiety symptoms (19%). Other targets included: general distress, stress and other outcomes	Fifty-three symptomatology measurement instruments were used in the included studies. The average number of instruments was 2.78 (SD = 1.31). Seven studies used the PHQ-8 (one study) or the PHQ-9 (six studies), seven used the GAD-7, five the DASS-21, 5 the BDI-II, 4 the PANAS, three used the Problem Distress Scale, 3 used a version of the PSS. Other instruments included the Problem resolution scale (2 studies), the HADS (2 studies), the QIDS-SR (2 studies), the Flourishing Scale (2 studies)

CBT Cognitive behavioral therapy; *MOL* Method of levels; *PST* Problem solving therapy; *VRET* Virtual reality exposure therapy; *AUDIT-C* Alcohol Use Disorders Identification Test; *BDI-I* Beck Depression Inventory-II; *BFI* Big Five Inventory; *BSC* Beck Stress Checklist; *CES-D* Center for Epidemiologic Studies Depression Scale; *DASS* Depression Anxiety Stress Scales; *DAST-10* Drug Abuse Screening Test; *EORTC QOLINFO2* European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire; *FS* Flourishing Scale; *GAD* Generalized Anxiety Disorder; *HADS* Hospital Anxiety and Depression Scale; *IDS-C* Inventory of Depressive Symptomatology- Clinician Rated; *MBI* Maslach Burnout Inventory; *MMSE* Mini-Mental State Examination; *PANAS* Positive and Negative Affect Schedule; *PHQ* Patient Health Questionnaire; *PSS* Perceived Stress Scale); *QIDS-SR* Quick Inventory of Depressive Symptomatology Self Report; *SF-36* Short Form Health Survey; *SIGH-D* Schedule for Affective Disorders and Schizophrenia—Depressive section; *SWLS* Satisfaction With Life Scale; *WHO-5* World Health Organization Well-Being Index

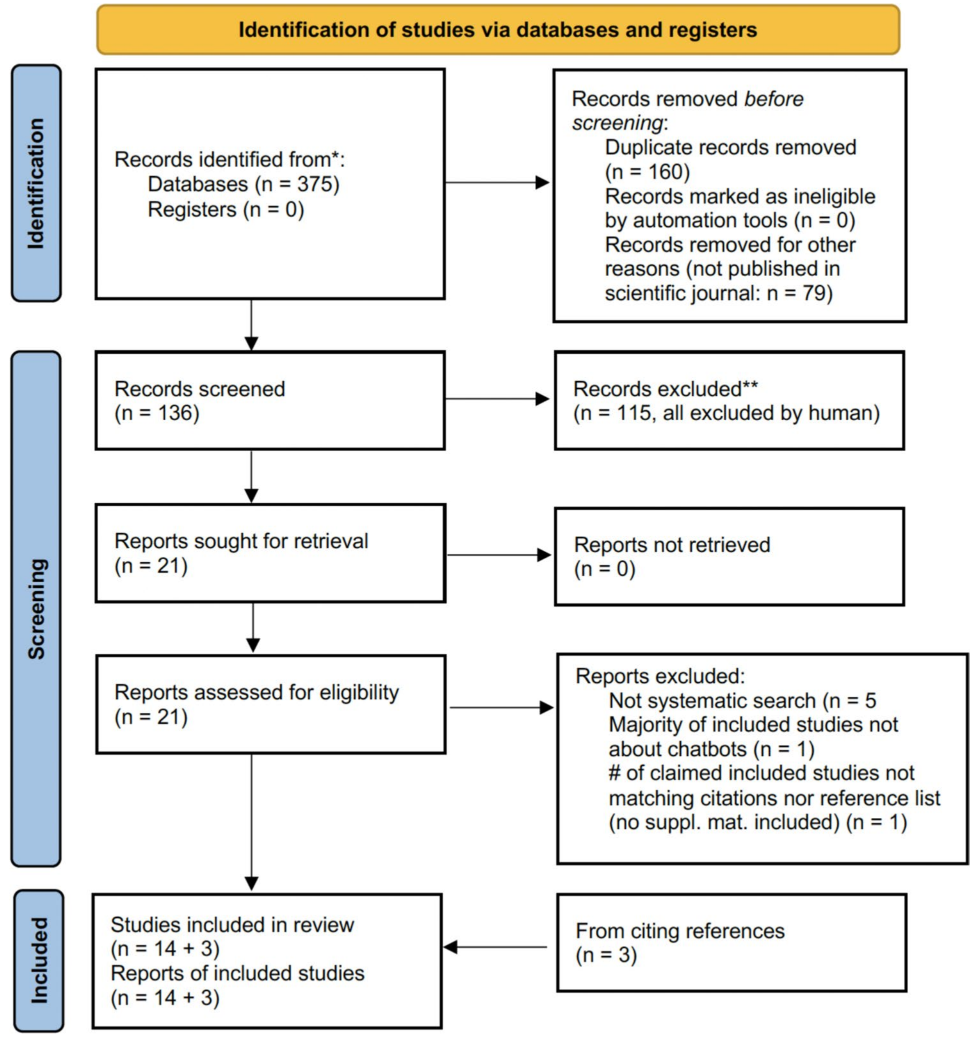
Table 4 Chatbot purposes/roles, embodiment and modes of interaction

	Chatbot purposes and/or roles	Percent embodied	Mode of interaction, dialogue initiation
Abd-Alrazaq et al. (2021):	"therapeutic purposes (n=12), training (n=9), self-management (n=6), counseling (n=5), screening (n=4), and diagnosis (n=1)" (p.7)	82%	Unclear
Abd-Alrazaq et al. (2019):	therapy and counseling (17+5 studies), training and education (12+4 studies), self-management (4 studies), diagnosis (2 studies)	83%	Users could only use the mouse and keyboard to interact with chatbots in 49% of the studies (as opposed to voice, or voice and text), whereas chatbots interacted vocally with users in 55% of the studies
Ahmed et al. (2023):	"diagnosis" (assessment): 24%, therapy and counseling (21%+12%), education (17%), others: mostly combination of the above; Care receiver in 7% of studies	29%	In 48% of the studies, dialogue was initiated by the chatbot, in 36% by the user, and the remainder was either mixed or undisclosed in the studies. In the majority of the studies, the interaction with the chatbot was performed through textual input (71%)
Bendig et al. (2022):	symptomatology reduction	Not described	Not described
Jabir et al. (2023):	Goal. treatment and monitoring: 66%; education and training: 34%; Role. coach: 72% of studies, healthcare professional: 19%, informal: 6%	44%	Unclear
Provoost et al. (2017):	Role. Social interaction partner: 23 studies; coach: 12 studies; tutor: 10 studies; healthcare provider: 10 studies	100% (as for research question)	Input: 11 studies included communication with speech, 9 with facial and gaze expressions, 9 with hand and body gestures, 5 with touch, 3 with text. Output: 34 studies included communication with speech, 30 with gaze and expressions, 25 hand and body gesture, 8 through text. (mismatching values for users and chatbots not explained; could be input options were more limited)
Martin and Richmond (2023):	“The focus of these conversational agents was varied and included stress, life skills coaching, mental health literacy, well-being, and social support for both children and their families." (p.7)	40%	Eight studies included a textual mode of interaction with the chatbots and 6 studies included a verbal interaction mode. One study included touch-screen and voice
Martínez-Miranda (2017):	Role. Counselor: 3 studies (aim: symptom reduction); Patient: 3 studies (aim: education)	100% (as for research question)	“responses are predefined” (p. 4) for one study, "responses are generated using inference rules” (p. 5) for another. Another example is that “responses match the student’s questions with a corpus of scripted dialogues in an internal database to select the most appropriate answer” (p. 7)
Vaidyam et al. (2019):	Five studies were used for monitoring, 3 for therapy and two for diagnostic	80%	Unclear. Primarily text, even if voice available
Vaidyam et al. (2020):	Therapeutic efficacy (3 studies); diagnostic (2 studies), acceptability (2 studies)	29%	The user input was constrained in four of the seven studies. Six chatbots had a primary mode of interaction through text, while voice was also available. One of the chatbots allowed only for vocal interaction
Gaffney et al. (2019):	symptom reduction (8 studies); wellbeing improvement (3 studies); self-management (2 studies)	54%	In four studies, user input consisted in choosing predetermined options, whereas the in the other studies it was written or spoken. The output consisted on displayed text in seven studies and spoken content in four (not disclosed in two studies)
Ogilvie et al. (2022):	alcohol education and risk assessment, drugs and opioid abuse monitoring, smart assistants (e.g., Alexa, Siri) as support agents for SUDs	None	The technical characteristics of chatbots were not systematically described. For example, the authors mention AIML was used to design the chatbot in a study, and that reddit was crawled to build a database of questions and answer in a different study, with the chatbot responses being generated on this basis and the users’ input using deep learning, and that smart assistants (e.g., Alexa, Siri) were used in another study
Otero-González et al. (2023):	detection	Unclear	The majority of studies implemented interaction with the chatbot through chat (21 studies), five did through voice and chat, five through voice, two through voice and video. Two more through voice, chat and video; and one through chat and video
Pacheco-Lorenzo et al. (2021):	detection	Unclear	Unclear

Table 4 (continued)

	Chatbot purposes and/or roles	Percent embodied	Mode of interaction, dialogue initiation
Abd-Alrazaq et al. (2020):	therapy (10 studies); self-management (2 studies)	50%	In 75% of the studies, conversation was controlled by the chatbot. In 75% of the studies, the user could only interact with the chatbot through text and point-and-click; Modes of input. Only written: 9 studies; Only spoken: 2 studies; Written and spoken: 1 study. Modes of output. Only written: 6 studies; Written, spoken, visual: 3 studies; Spoken, visual: 2 studies; Written, visual: 1 study
Lim et al. (2022)	therapy	27%	"Three trials had chatbots with response options and written as the input format. Three trials had chatbots that used a combination of written, spoken, and gestures as the output format." (p. 337)
He et al. (2023):	symptom reduction	28%	Unclear

Fig. 1 PRISMA flowchart


Identification of studies via databases and registers
Identification
Screening
Included
Records identified from*:
Databases (n = 375)
Registers (n = 0)
Records removed *before screening*:
Duplicate records removed (n = 160)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (not published in scientific journal: n = 79)
Records screened (n = 136)
Records excluded** (n = 115, all excluded by human)
Reports sought for retrieval (n = 21)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 21)
Reports excluded:
Not systematic search (n = 5
Majority of included studies not about chatbots (n = 1)
# of claimed included studies not matching citations nor reference list (no suppl. mat. included) (n = 1)
Studies included in review (n = 14 + 3)
Reports of included studies (n = 14 + 3)
From citing references (n = 3)

measures. The authors highlight the potential of chatbots to improve access to mental health care, notably in developing countries, where the need is particularly urgent.

The scoping review by Ahmed et al. (2023) also examined the features of chatbots, this time used for anxiety and depression specifically. The majority of the studies reported on chatbots implemented as standalone software (55%), followed by web-based (19%), the remainder were either both standalone and web-based or not reported. The authors report the chatbots included AI (69%) or rule-based (19%) components for response generation, while the rest employed both. Both depression and anxiety were targeted by 60% of the chatbots, 38% targeted only depression, and one study targeted anxiety alone. While 24% of the included

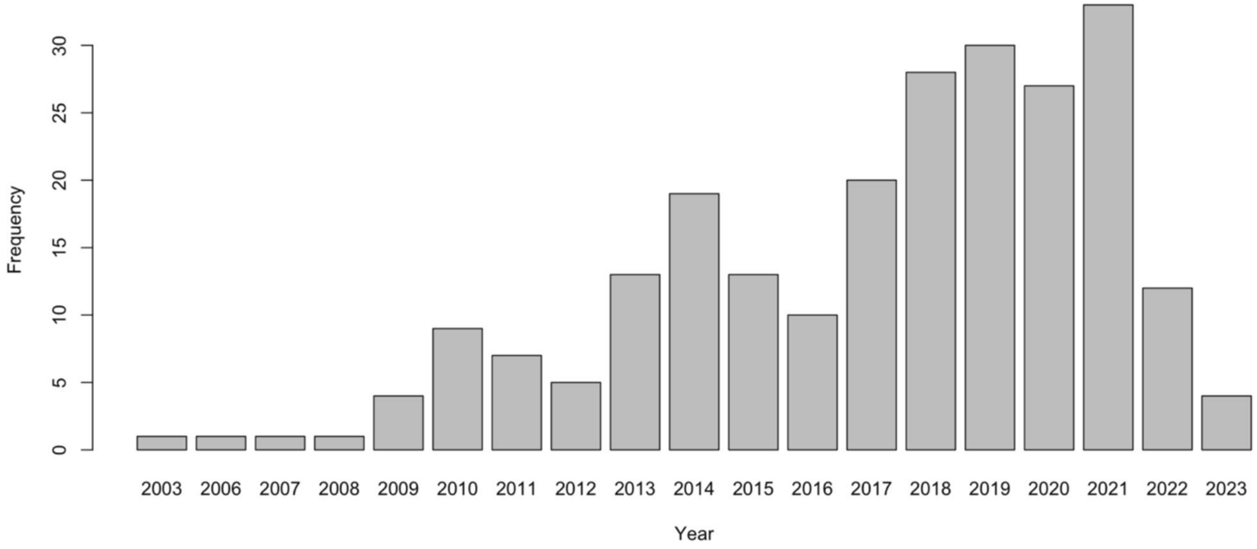


Fig. 2 Number of articles by year of publication

studies reported on user perceptions and engagement, the review did not report on these findings. The review mentions that effectiveness of chatbots are frequently reported in the included studies (but only nine studies, i.e., 21%, had pre-post measurements of outcomes). The review itself doesn't focus on effectiveness. But from the overall findings of the studies they assessed, they note that "chatbots are to be considered an effective tool for depression or anxiety" (p. 10). The authors mention that most chatbots in the included studies were in the early stages of evaluation. Interestingly, while many chatbots emulated care providers, others emulated patients, allowing users to gain understanding of their own situation caring for the chatbot.

The scoping review by Bendig et al. (2022) is interested in the scope and evidence of chatbots in promoting mental health, focusing on assessing the risks, opportunities, and challenges of such use. All chatbots aimed at symptomatology reduction. The durations of the interventions ranged from 20 min to 10 weeks. Half of the studies had an active control group, while the other half had a waitlist control group. Overall, of the studies that assessed multiple outcomes, most found at least one for which the chatbot intervention was superior to the control group, while one or more was not. In the studies examining only one outcome, the chatbot intervention was not found to be superior to the control group. Users perception and engagement were assessed in several studies with ratings indicated good satisfaction with the interventions. However, the authors note that not only effectiveness and users' perceptions, but also privacy and data protection are important issues to address in the provisioning of chatbots for mental health.

The scoping review by Jabir et al. (2023) is interested in the outcomes targeted studies using chatbots as interventions and the measures used to assess them. The duration of the intervention ranged from one day to nine weeks. The chatbot was proposed in a standalone application on computer, tablet or smartphone in 57% of the studies. It was a web-based application in 22% of studies, messaging-based (e.g., SMS) in 19%, and mixed in 4%. Dialogue generation was rule-based in 56% of the studies and relied upon AI in 44%. The review reports an overall total of 203 distinct outcome measures, of which 150 were designed by the authors of the studies. Validity evidence for about half of these instruments was lacking. The number of outcomes assessed in a single study ranged from two to 17. Most outcomes were clinical, followed by user experience outcomes. Two studies assessed technical outcomes, and one study examined other types of outcomes. The review doesn't systematically mention findings of the included studies but reports that better user perceptions were associated with greater improvements in outcomes. Among the studies with a comparator group (21 in total), 67% found higher effectiveness of the chatbot intervention compared with control, while 25% found no difference between the groups, and 19% reported mixed evidence. The authors observed that outcomes were frequently measured using pen and paper (with more than half of the studies assessing five or more outcomes through this method). They also noted that chatbots or smartphone sensing could be used to alleviate participants' burden in filling out questionnaires.

The scoping review by Provoost et al. (2017) is interested specifically on embodied conversation agents for promoting mental health from a technical perspective. Fourteen studies used a web-based chatbot, 12 studies incorporated the chatbot in the context of a serious game, 10 studies used standalone software, and three studies used virtual reality

(VR). Among autism spectrum disorder interventions, many focused on training joint attention skills and imitation skills and other communication skills in children. Other applications included the development of skills for job interviews. The authors mention that sample sizes in this area are typically very small, but the results were promising. However, the authors do not specify the instruments used to measure the outcomes. The interventions targeting depression oftentimes aimed at increasing self-disclosure from participants. The review suggests that chatbots for symptom alleviation appear promising, particularly when they include empathic expressions, and that overall evidence hints at good acceptability and engagement with chatbots for mental health. However, because the chatbots in the included studies were in an early stage only (e.g., development phase or pilot studies), the authors consider there was insufficient evidence to suggest chatbots can be used as alternatives to existing interventions or as effective adjuncts to treatments. The authors note efforts have been made in some of the included studies in relation to the personalization of chatbots.

The scoping review by Martin and Richmond (2023) is focused on chatbots used for promoting mental health in children and their caregivers. The review mentions that the included studies did not focus on examining the effectiveness of the intervention with regards to clinical outcomes. Indeed, measures of clinical outcomes are mentioned for only three studies (20%). The summaries of the results of studies in Table 4 of Martin and Richmond (2023) confirm this, yet shows in one study social anxiety of children was reduced by chatbot interaction to the same extent as through behavioral treatment, that in another study participants found the chatbot helpful, and that in another study, participants found the intervention of low quality, despite enjoying interactions with the chatbot. The overall findings suggest that participants were satisfied with the chatbots, rated the intervention high in usability, and found it acceptable. While chatbots for children may be currently adequately designed, evidence for their effectiveness is still lacking. The authors mention that there are very few interventions specifically targeted at the mental health of parents, which might be needed given the specific stressors involved in parenting.

Unlabeled reviews

The review by Martínez-Miranda (2017) is interested in the use of embodied conversational agents (chatbots with an animated representation) for detecting and preventing suicidal behavior. Three publications featured chatbots acting as counsellors. The author notes that none of these studies focused directly on suicidal behavior. Given the focus of the review, all chatbots in the included studies were embodied. In one study focusing on depression, suicidal ideation and behaviour were identified through interactions with the chatbot system. In two other studies focusing on depression, suicidal ideation was assessed using the Patient Health Questionnaire – 9 (PHQ-9, item 9), rather than through interaction with the chatbot. The chatbots in three publications (six studies) acted as patients. One study used a chatbot acting as a patient with bipolar disorder to teach medical students how to assess suicide risk. The students queried the chatbot for information or learned from watching a video (the control group). Chatbot users asked more suicide risk and bipolar disorder relevant questions compared to the control group participants. Lastly, two publications focused on teaching the identification of suicide risks and other risks in diverse populations (four studies) for the first and in American Indian and Alaskan community gatekeepers for the other using the same chatbot. Communication with the chatbot involved both verbal and non-verbal communication (e.g., facial expressions). The author mentions that none of the included studies explicitly targeted suicide risk and that "an explicit mechanism to detect and cope with suicide risk is not present in all existent works" (p. 9).

The review by Vaidyam et al. (2019) is interested in the use of chatbots in mental health. The review does not systematically describe the theoretical framework of the included studies, nor the countries from which the studies originated. The review mentions the included studies show promise for the use of chatbots to reduce mental health symptomatology (depressive symptomatology particularly) and increased adherence, with results from several studies supporting these claims. The review also mentions a preference for interacting with chatbots over clinicians in three studies. The reported studies found a high participant satisfaction across the studies. The authors mention main benefits of chatbots are psychoeducation, adherence and cost, as well as the importance of alliance building. The review also mentions there is limited potential harms from interactions with chatbots, but notes the inability to respond to crisis situations, exaggerated attachment to the chatbot, and data breach as potential risks. The review emphasizes the lack of research on the best chatbot modalities in terms of adherence and engagement.

Systematic reviews

The systematic review by Vaidyam et al. (2020) is interested in updating the review by Vaidyam et al. (2019) regarding "*psychiatric chatbots*. The authors queried four databases in January 2020, limiting to studies published after July 2018. The duration of the intervention lasted between three hours and eight weeks. The main clinical targets were depression (4 studies) and anxiety (3 studies). Other targets included wellbeing (2 studies), broad spectrum of mental disorders (assessment). Perceptions of chatbots and adherence, assessment of emotions were also investigated. The theoretical frameworks for the interventions is not reported. Three studies (one of them with a control group) assessed the efficacy in symptom reduction in reported improvements in mood, psychological distress and general wellbeing, overall reporting success at symptom reduction, particularly in users more engaged with the chatbot. Two studies assessed agreement in diagnosis between chatbot and therapists; one showed good agreement, while the other found moderate agreement. The two studies assessing user perceptions found users were in majority adherent and engaged.

The systematic review by Gaffney et al. (2019) is interested in the effectiveness of chatbots in alleviating mental health issues. The chatbots in five publications were web-based. The chatbots were standalone on a computer in four studies, in a smartphone app in three studies and in VR in one study. The authors report important variations in the engagement with the chatbots between studies.

The principles included in the chatbots were inspired by CBT in four studies, method of level in two studies, and mindfulness in one study. One study implemented communication enhancement, while four others integrated multiple therapeutic approaches.

The comparator groups were treatment as usual in two studies, waitlist in one study, the Eliza agent in two studies, information (e-book or CDs/mp3s) in three studies, an attention control in one study, users who didn't use the downloaded app in one study (quasi-experiment), and no intervention in one study. The authors report that in five studies with a control group, the chatbot intervention improved symptomatology better than the control condition, but in four studies with a control group, the reduction was similar between groups. Overall, studies reported high participant satisfaction. For instance, two studies noted that participants would recommend the chatbot over a clinician.

However, several challenges to the use of chatbots in the treatment of mental health symptomatology were reported. The most frequently reported issues were: The users find the content repetitive in six studies, there are issues with the appropriateness of the generated responses to users' queries (chatbot's "understanding") in four studies. Less frequent concerns included: low quality of voice of chatbots, superficial relationship with chatbots, the lack of delivery of specific tools and interactions with the chatbot not long/frequent enough according to some users.

The systematic review by Ogilvie et al. (2022) is interested in chatbots as support tools for individuals with substance use disorders (SUDs). The theoretical frameworks underlying the interventions were not specified. Among three studies using the Alcohol Use Disorders Identification Test – Consumption, one did not test the effectiveness of the chatbot. One study (no control group) showed improvements in substance use, cravings and confidence, as well as depression and anxiety. The RCT showed significant decrease in substance use compared with the control group, but no difference in depression and anxiety.

Studies examining user experience and perception found that while users were generally open to using chatbots, they also expressed concerns about the lack of empathy of chatbots. Other areas of concern where responses that were too verbose, or an unappealing user interface. Another study used a multiple-choice response format in relation to questions about addiction and substance use; the findings of the study are not clearly reported in the review.

The systematic review by Otero-González et al. (2024) focuses on the use of chatbots for the screening of depression. The chatbots we designed as to offer interactive implementations of clinical assessments. The authors note that 23 chatbots included "emotion modules". Ten studies at least employed readily available frameworks for the creation of chatbots featuring natural language processing (NLP; DialogFlow, RASA, Aria Valuspa). Two studies used the Tess chatbot by X2AI. One study created an Alexa skill for its chatbot. Python was mentioned as the programming language for two studies with no further information on the technologies used. Twelve other studies were reported to use a self-developed chatbot without further information. One study used Alexa, Siri, Cortana and Google Assistant. Fifteen studies relied upon a smartphone app, three studies a computer app, two studies a web app. This information is not reported for other studies. Whereas the review mentions depression instruments, notably, were used to examine the validity of the screening of depression through chatbots (e.g., Table 5 in Otero-González et al. (2024) presents the instruments, which are summarized in Table 3 of the present review), it does not explicitly mention this was actually achieved – i.e., a good correspondence of the chatbot assessment with that of the instruments, or in how many studies this was the case. The authors outline a key limitation in the field: The pace of technological advancement outpace the timeline required for conducting multi-stage trials. This could explain the lack of such projects in the field. They also mention the importance of the ability of the chatbots to relate to the users and form relationships with them as a requirement for their success.

The systematic review by Pacheco-Lorenzo et al. (2021) is interested in the use of chatbots for the early detection of psychiatric disorders and neurocognitive disorders. Eight studies used samples composed of elderly people, while the others didn't have restrictions based on age (age distribution not specified). A notable aspect of this review is that most studies gathered data directly through chatbots, although some relied on alternative sources. Eleven studies used visual and audio features for detection, nine studies used language and three studies were focused on assessing the chatbot rather than its assessment capabilities.

In terms of technological infrastructure, six studies used a machine learning toolkit (e.g., Apache OpenNLP, Keras, Scikit-learn, OpenFace). Two used frameworks specialized in building chatbots (e.g., DialogFlow, Aria Valuspa). Additionally, two studies relied on pre-existing chatbot systems (RoBoHon, Monarca). Four chatbots were proposed on Android, three on Windows, two on iOS, and two used Google Cloud (web-based). The authors mentions good correspondence between the use of chatbot-elicited data and the validation instruments used for the detection of the targeted disorders. However, the authors also point out certain limitations in the performed validity assessments.

Meta-analyses

The systematic review and meta-analysis by Abd-Alrazaq et al. (2020) is interested in obtaining pooled estimates of effect sizes regarding the reduction of mental disorders symptomatology following the use of chatbots for mental health and in summarizing the literature. The authors noted: “Chatbot responses were based on predefined rules or decision trees (rule-based) in two-thirds of studies (8/12). Chatbots in the remaining one-third of studies (4/12) utilized machine learning and NLP to understand users’ replies and generate responses.” (p. 7). Half of the chatbots were web-based and half ran on standalone applications. The follow-up periods ranged from two weeks to 12 weeks. While a meta-analysis was conducted specifically for depression and anxiety, the other outcomes were narratively summarized. Chatbots were more effective than treatment as usual or information for reduction in depressive symptomatology. For the reduction in symptoms of anxiety, the authors note no difference between chatbot use and receiving information. Of the two studies using the Positive and Negative Affect Schedule (PANAS) scales as outcomes, only one demonstrated greater effectiveness for chatbots compared to information provision. Subjective wellbeing was improved in one of three studies. The authors further indicate chatbot use seems to be efficient at decreasing psychological distress compared with no intervention and at reducing stress and acrophobia (better than therapist-lead intervention in this last instance).

The systematic review and meta-analysis by Lim et al. (2022) is interested in the effectiveness of chatbots targeting depression and anxiety and focuses on RCTs. The control groups were wait-list (5 publications), treatment as usual (4), and information only (2). Three studies used embodied chatbots. Six publications used the Deprexis intervention, while the remaining each used a different chatbot. In eight studies, the intervention was provided online. The meta-analysis centered on depression, and results indicate overall that the interventions were effective ($g = 0.54$, medium effect size). Subgroup analyses didn’t reveal significant differences between subgroups, although the effect sizes in the comparisons sometimes varied considerably (e.g., PST intervention only, $g = 1.05$; for mixed psychotherapy, $g = 0.51$).

The systematic review and meta-analysis by He et al. (2023) is interested in the effectiveness of RCTs involving chatbots for mental health. Sixteen chatbots provided emotional and empathic responses, while 11 studies automated reminders to interact with the chatbot were sent to participants. The duration of the interventions varied considerably, with about as many interventions lasting zero to four weeks, five to eight week or more than nine weeks. A majority of the studies (53% had no follow up), the remainder were fairly evenly divided between zero to eight-week follow ups and more than nine-week follow ups. The authors report results separately for each of the clinical targets. For depressive symptoms, effect sizes compared with control were small short-term and smaller long-term. In the case of generalized anxiety symptoms, the authors report a small effect size short-term. These effects vanished at long-term follow-up. For specific anxiety symptoms, the authors report significant effects in the main analyses (small to moderate effect size) short term and a non-significant long-term efficacy. The pooled estimate for quality of life or wellbeing was small short term but non-significant in the long-term. For General distress, small effect sizes were observed for short-term and for long-term efficacy. For the outcome stress, the authors report small effect sizes short-term, which vanish in the long-term.

Among numerous moderators reported by the authors, empathic responses were linked with higher efficacy for all outcomes. Dose–response was significant for outcomes depression and generalized anxiety symptoms but not for other outcomes. Gender was significant for generalized and specific anxiety symptoms but not the other outcomes. The effect sizes were larger when the intervention targeted the outcome directly compared with not directly for depression symptoms, general anxiety symptoms, general distress and stress. Other included outcomes were mental disorder symptoms (small effects short-term and long-term), psychosomatic disease symptoms (medium effect size short-term and small effect size long-term), positive affect (non-significant) and negative affect (small effect sizes short-term, non-significant long-term). However, when chatbot interventions were compared with active controls, only depression and general distress were better improved by chatbots. The authors mentions favorable user perceptions and adherence in most studies reporting on such aspects.

Synthesis

This scoping review reports on 17 existing reviews examining the use of chatbots for mental health. These reviews included six to 54 articles. Interestingly, more of the studies included in the reviews were published from 2018 (134 studies) than before (104 studies) and most of the studies were included in only one review (72%). This can be notably explained by the fact that the focus of the reviews were partly diverging, but also by differences in search strategies and screening processes. Further, some studies were published after the literature search was carried out for some reviews, and some later reviews only included recent studies (e.g., Ahmed et al., 2023).

The reviews showed considerable variation in thematic focus. Some explored chatbot features (Ahmed et al., 2023), user perceptions (Abd-Alrazaq et al., 2021), outcome types and measurement strategies (Jabir et al., 2023), and chatbot efficacy (He et al., 2023). Others were more specific in scope, focusing on either general mental health (e.g., Bendig et al., 2022), particular disorders (e.g., Ogilvie et al., 2022), or specific populations like children and adolescents (e.g., Martin & Richmond, 2023). Among the 17 reviews, seven were scoping reviews, eight were systematic reviews (three of which included meta-analyses), and two were unlabelled reviews.

Ten reviews reported, either in the article or the supplementary materials, the countries of origin for included studies, with the USA being the most frequent, followed by the UK, Australia, and Japan. The proportion of studies with clinical samples varied importantly between the reviews (e.g., one study in Bendig et al., 2022, 62% in Gaffney et al., 2019; five reviews did not provide such information). The majority of the reviews mentioned less than 15 participants in one or more of the included studies. On the contrary, some of the studies had rather large sample sizes (up to 32,096 participants in one specific study). The majority of participants were adults, although the proportion of children and adolescents was high in two reviews (Martin & Richmond, 2023; Provoost et al., 2017). Gender representation varied widely across studies. Some included 100% female participants, while others involved mostly men (up to 93%). For instance, Abd-Alrazaq et al. (2019) reported a slight majority of male participants while He et al. (2023) report women are more represented. The gender composition of the samples in the included studies is not reported upon in several reviews.

Technological characteristics of the chatbots also varied. Most were implemented in stand-alone computer applications, followed by web-based, and less ran on smartphone apps. Chatbots were embodied in a majority of studies, but maybe less so in the most recent reviews. The use of VR seems emergent in the field. Many chatbots were rule-based, others relied upon AI, while a few used both approaches. Some reviews mention most chatbots to be rule-based (e.g., Abd-Alrazaq et al., 2019), while others mention most rely upon artificial intelligence (e.g., Ahmed et al., 2023). This might signal an evolution in more recent studies. A related evolution might be the use of tools specialized in building chatbots rather than designing them from the ground up, as several recent reviews mention the use of such technology (e.g., Otero-González et al., 2024; Pacheco-Lorenzo et al., 2021). Textual interaction appeared to be the most frequent mode, with less chatbots relying primarily on voice. While personalization is a feature of some of the chatbots in the review by Provoost et al. (2017), it is unclear how such possibility affects users' experience and the effectiveness of chatbots.

The majority of the reviews mentioned principles of CBT were included in the design of chatbots in some of the included studies. The purpose of the chatbots varied and included: the emulation of a therapist, counselor, or coach by providing self-management strategies and life skills), of an interaction partner. Less commonly, some chatbots were designed to emulate a patient, which the users could gain understanding of their own situation and of others while caring for the chatbot (e.g., Ahmed et al., 2023; Martínez-Miranda, 2017).

The primary goal of most chatbots was symptom reduction, while others were aimed at monitoring and assessment or training. Of the reviews which didn't have a restricted focus regarding the clinical targets of the included studies, the most frequent targets appear to be depression, anxiety (as well as autism for reviews which didn't restrict on age). General mental health and subjective wellbeing also appear to be frequent targets. Overall, the PHQ (versions with 9, 8, or 2 items) was the most frequently used clinical followed by the Generalized Anxiety Disorder 7 (GAD-7) scale, the PANAS, and the Beck Depression Inventory-II. Jabir et al. (2023) noted that many studies still relied on pen-and-paper to collect data on outcomes might be burdensome to participants. They recommended incorporating chatbot-based measures or smartphone sensing to alleviate this burden.

Two reviews focused specifically on the use of chatbots for the detection of symptomatology (Otero-González et al., 2024; depression) and Pacheco-Lorenzo et al. (2021; mental and neurological disorders), highlighting the possibilities for such assessements, but Vaidyam et al. (2020) report less optimistic results from one study.

The meta-analyses (Abd-Alrazaq et al., 2020; He et al., 2023; Lim et al., 2022) reported unequivocally improved depressive symptomatology. The meta-analysis of Abd-Alrazaq et al. (2020) didn't find significant pooled effects for anxiety, while the meta-analysis of He et al. (2023) found significant overall pooled effects for general and specific anxiety symptoms, as well as quality of life, general distress and stress among others. In comparisons with active control groups, chatbots only had an advantage for the reduction of depressive symptomatology and general distress. He et al. (2023) also identified several significant moderators of the effectiveness of chatbots. The other reviews were generally less focused on effectiveness and provided sometimes mixed evidence. The reviews overall found most studies reported positive users' evaluations, for instance in terms of usability, acceptability, engagement and adoption. Nonetheless some concerns have also sometimes been raised. Users sometimes reported repetitive or scripted responses, a lack of understanding, insufficient empathy, and reduced attractiveness, particularly for embodied chatbots (e.g., Abd-Alrazaq et al., 2021; Gaffney et al., 2019).

The question of safety, privacy and data protection has been raised by Bendig et al. (2022). At this stage, the reviews mention very little, if anything, on this respect. While web-based chatbots feature better accessibility and do not require application installation (Abd-Alrazaq et al., 2019), privacy is probably better preserved with standalone software. Abd-Alrazaq et al. (2020) and Vaidyam et al. (2019) reported that chatbots have minimal reported risks in the studies they examined, while Ogilvie et al. (2022) appear more circumspect, particularly in the use of smart assistants (generic purpose, e.g., Alexa, Siri) as supportive agents.

Provoost et al. (2017) noted that studies of chatbots (included in their review) were generally at very early stages and considered this precluded the recommendation of the use of chatbots over human therapists. Accordingly, Martin and Richmond (2023) classified 87 percent of the included studies in the category preliminary testing. They noted "The field mainly consists of preliminary design studies, with a large focus on evaluating feasibility and acceptability over efficacy." (p. 11). Similarly, Pacheco-Lorenzo et al. (2021) highlighted the frequent lack of validation in the studies they examined and the predominance of small sample sizes in most of these studies. Ogilvie et al. (2022) noted the lack of control active comparator groups in the included studies, while Provoost et al. (2017) and Vaidyam (2019) highlighted the need for comparisons of the chatbots with other forms of digital interventions as well as the lack of study of long term benefits of chatbots. Provoost et al. (2017) also mentioned the lack of comparison of chatbots effectiveness with conventional treatment. Vaidyam (2019) mentioned the possibility of over-reliance upon chatbots which might negatively impact users. Gaffney et al. (2019) noted chatbots might be unable to respond appropriately in high risk situations (e.g., suicidal ideation). With the exception of the meta-analysis by He et al. (2023) and the systematic reviews by Bendig et al. (2022) and Gaffney et al. (2019), the reviews reported on few RCTs, and many included studies indeed did not feature a comparator group. The sample sizes of most studies included in the reviews (but not all) were too small to be properly powered as the effect sizes were overall generally small for most outcomes (He et al., 2023). Indeed, Otero-Gonzalez et al. (2024) found a significant proportion of studies had small sample sizes. Pacheco-Lorenzo et al. (2021) noted the frequent lack of validation in the studies they examined and observed small sample sizes in most of these studies.

Some reviews noted that the question of the risk of bias and variability in reporting could be a concern in the field. Notably, Abd-Alrazaq et al. (2020) noted risk of bias could have affected the results of several of the studies they included in their review. He et al. (2023) reported that only 25% of included studies had a low risk of bias, whereas 40% had a high risk of bias. Similarly, several studies were reported to have high risk of bias in the review by Lim et al. (2022) and that performance bias might have influenced results of several studies due to lack of participant blinding. Gaffney et al. (2019) similarly observed high variability in study quality and emphasized the absence of long-term follow-ups, which limits understanding of the sustainability of chatbot benefits. Additionally, Vaidyam (2019) noted considerable variation in reporting and use of measures in the included studies which complicate comparisons between studies.

Discussion

This scoping review of reviews on chatbots for mental health, based on a systematic search across four databases, identified 17 eligible publications. We have provided summaries of these reviews and an overarching synthesis.

Answering the research questions

Is the field of chatbots for mental health growing?

There has been a recent increase in the number of studies in the field in the recent years. Over half of the studies included in the reviews were published in 2018 or later. This might be due to diminished technical barriers, as the recent years have seen the apparition of ready-made frameworks for building chatbots (e.g., DialogFlow, Rasa).

What are the research foci of the reviews in the field?

Overall, the reviews focused on the reduction of mental disorder symptomatology and other clinical targets (e.g., stress, wellbeing). Other aspects were examined in the reviews, such as the use of chatbots for the detection or assessment of mental disorder symptomatology, the features of the chatbots and users' perceptions and engagement.

Which are the countries from which most studies in the field originate?

It was apparent that most studies included in the reviews originated from countries in which English is the (or an) official language. More studies originating from other countries could permit determining whether users' perceptions and the efficacy of chatbots is universal, or rather dependent upon cultural parameters.

What are the sample sizes of the studies reported in the reviews?

As we have mentioned above, several studies had large samples. But sample sizes in many studies were probably not sufficient, given the expected effect sizes for such studies (He et al., 2023). In particular, Provoost et al. (2017) reported a mean sample size lower than 20 among studies focusing on autism spectrum disorder and psychosis. Further most reviews included one or more studies with 15 participants or less.

What are frequent characteristics of the interventions (chatbots)?

Embodiment and textual interactions are frequent characteristic of chatbots. The use of voice was a less frequent feature, particularly on the users' side for interaction. Interestingly, chatbots are programmed to play different roles, such as therapists, coaches, fellow patients or simply social partners. Treatment and monitoring as well as education and training were frequent goals of the chatbots. It was not clear in the reviews what proportions of chatbots were designed for maintaining a relationship with the user over extended periods of time (relational chatbots; see Bickmore & Gruber, 2010). It can be postulated that for chatbots to exhibit similar efficacy as therapists, such focus should be primordial. Further, Gaffney et al. (2019) report that "participants valued aspects of agents usually seen as unique to therapy with a human, such as empathic responses, personality, the ability to build a relationship, and an interactive, conversational approach.)" (p. 7).

What is the degree of scientific rigor (e.g., use of RCTs, adequately sampled studies) in the field?

It was apparent from the reviews which didn't focus exclusively on RCTs (e.g., He et al., 2023), that properly sampled RCTs were a minority in the field (Abd Alrazaq et al., 2019). This is problematic as only properly sampled RCTs (ideally with an active control group) permit to truly assess the potential of chatbots for mental health. Many studies lacked any comparator group, making it difficult to distinguish between chatbot effectiveness and natural symptom remission.

What are the main clinical targets of the studies and the most frequent measures?

Most studies included in the reviews which didn't focus on specific disorders targeted depression, anxiety as well as general mental health / wellbeing. Autism spectrum disorders were the most frequent clinical targets in research involving children. The PHQ-9, the GAD-7 and the the Depression Anxiety Stress Scales 21were the most frequently used measures due to the large focus of the studies on depression and anxiety. While the large majority of studies collected data using questionnaire instruments, a minority used clinical interviews.

Are the interventions efficient at improving mental disorder symptomatology and other targets?

The meta-analyses of Abd-Alrazaq et al. (2020), Lim et al. (2022), and He et al. (2023) provide the most robust evidence for answering this question. Focusing on depression, Lim et al. (2022) found medium effect sizes, attesting of the efficacy of chatbots for the reduction of depressive symptomatology. While Abd-Alrazaq et al. (2020) found only depression was improved by interaction with chatbots for mental health, but not anxiety. He et al. (2023) found that studies with passive control groups showed significant improvements across all clinical targets, but that only depression and general distress improved significantly in studies with active control groups. This discrepancy may be attributed to small sample sizes contributing to greater heterogeneity, or the inclusion of therapist-led interventions in the active control category, potentially minimizing observed differences in effectiveness. Such findings may, in fact, support the potential of chatbots as viable alternatives to traditional therapy in certain contexts. This issue could be investigated in priority in further studies. The use of chatbots seems to be efficient at alleviating at least depressive symptomatology (which is the largest contributor to disability among all causes) as well as general distress (He et al., 2023). Yet, users might overestimate what a chatbot can and cannot do (therapeutic misconceptions, see Khawaja & Bélisle-Pipon, 2023).

What are users' perceptions about chatbots for mental health?

Reviews addressing user perceptions, such as Abd-Alrazaq et al. (2021), reported generally positive users' perception (e.g., satisfaction, usability, acceptability) and generally a good engagement with the interventions. However, users may overestimate what a chatbot can do (therapeutic misconceptions (Khawaja & Bélisle-Pipon, 2023). It is important for developers of chatbots for mental health to clearly communicate the limitations of these tools to users. Additionally, some users expressed concerns about the lack of empathy in chatbot interactions and the repetitiveness of the responses.

Are there evolutions of the field that can be anticipated?

The included reviews provide partial insights into this question. Earlier reviews (e.g., Abd-Alrazaq et al., 2019) found response generation to be more frequently rule-based while Ahmed et al. (2023) found most chatbots used AI. This might signal an evolution in more recent studies. Indeed, frameworks allowing the rapid development of chatbots using NLP, offering more natural interactions compared to older technologies like AIML. These advancements may have contributed to the growing presence of mental health chatbots. It is probable this will continue to be the case in the future. For chatbots to be most useful, the systems should include a process allowing the assessment of the mental health of users and of their progress. However, many studies still rely on traditional pen-and-paper methods for outcome measurement (Jabir et al., 2023). Reviews by Otero-González et al. (2024) and Pacheco-Lorenzo et al. (2021) suggest that automated assessments via chatbot interactions are technically feasible, and this development appears to be a necessary evolution of the field. A promising direction for development of mental health chatbots involves enhancing their anthropomorphic features. Anthropomorphism refers to the ascription by people of human-like features (e.g., consciousness, intentions) to non-human entities (Seeger et al., 2021). In the context of interactions with chatbots and robots, it can be considered that higher anthropomorphism leads to more meaningful interactions with such entities leading to higher engagement (Wang et al., 2023). Although the included reviews seldom addressed anthropomorphism directly (e.g., Bendig et al., 2022), most of them mentioned design features that are typically aimed at increasing anthropomorphism (Herbener et al., 2024). Embodiment, for instance, was mentioned by the large majority of the included reviews. Martínez-Miranda (2017) and Provoost et al. (2017), specifically investigated embodied chatbots. The reviews also mentioned evidence that users of chatbots for mental health treated them as imbued to some extent with human qualities. Abd-Alrazaq et al., (2019) reported that chatbots were often considered friendly by users and at times emotionally responsive. Abd-Alrazaq et al. (2019) and He et al. (2023) also mentioned results indicating empathetic chatbots were generally more effective or preferred to those lacking empathy. Whether users believed chatbots were acting based upon intentions or thoughts has not been reported in the reviews. Interestingly, despite the fact that chatbots are not capable of intention or understanding, some users rely on chatbots for mental health as a means to test their own beliefs and emotional responses to situations (Grodniewicz & Hohol, 2024). Collectively, these findings suggest that increasing anthropomorphic features might be a promising avenue for further studies.

Implications

Several factors have been identified as areas for improvement in mental health chatbots, according to the authors of the included reviews. Despite good users' perceptions of chatbot for mental health (e.g., Abd-Alrazaq et al., 2021), the question of the lacking empathy and adaptability of chatbots to their users, and more generally the ability to respond in more relevant and natural ways have also been noted as potential deterrents to the use of mental health chatbots (e.g., Abd-Alrazaq et al., 2021; Gaffney et al., 2019; Oglivie et al., 2022). Other important aspects such as the detail and precision of the feedback from chatbots, along with their responsiveness and ability to foster a sense of being in users (e.g., Abd-Alrazaq et al., 2021; Gaffney et al., 2019; He et al., 2023; Oglivie et al., 2022). Users value continuous (daily) interaction with the chatbot, along with weekly summaries. They also tend to prefer embodied to non-embodied chatbots (but might disclose more to non-embodied chatbots). Vaidyam et al. (2019) noted the lack of research on the best choice of modalities in the dialogue of chatbots and users and reports arguments for voice (e.g., more natural) and text (e.g., discretion). Automatic reminders to interact can improve engagement but appear to moderate the effectiveness of interventions differently depending on the clinical target (He et al., 2023). More research is needed on this respect. The conversational abilities of chatbots also need to be improved (Abd-Alrazaq et al., 2021; Gaffney et al., 2019). He et al. (2023) noted: "The results of this meta-analysis demonstrate how the use of personalization and empathic responses can significantly improve the efficacy of CAIs [conversational agents]. In particular, empathic responses were linked to larger effect sizes for all mental health outcomes. This suggests that future technology and mechanism research on CAs may concentrate on these 2 capacities. A breakthrough in these 2 capacities will be necessary for CAs to function as competently as human therapists." (p. 19). But whether personalization, as afforded in some chatbots (Provoost et al., 2017), is associated with effectiveness and is a necessary avenue for future research (He et al., 2023). Kocaballi et al. (2019) describe two types of personalization in healthcare chatbots: implicit and explicit: "In implicit personalization, information needed for user models is obtained automatically through the analysis of observed user activities and interactions with the system. In explicit personalization, information needed for user models requires users' active participation in obtaining the required information." (p. 2).

Another important aspect relates to the assessment of the effectiveness of chatbots in the alleviation of psychiatric symptomatology, the improvement of wellbeing and the assessment and monitoring of users' status. Whereas the reviews indicate validated instruments are used in most of the studies, the review of Jabir et al. (2023) suggests about half of the used instruments are not validated. Additionally, the authors note that 83.7% of outcomes are self-reported by participants, which can place unnecessary burden on users. This issue could potentially be addressed by integrating assessment components directly within the chatbot programming, using successfully validated algorithms (Pacheco-Lorenzo et al., 2021). Importantly, many studies in the included reviews focused mainly on user's experience did not seem to evaluate the effectiveness of the chatbots. As a result, it remains unclear whether the intended goals were achieved or not in these instances. This is particularly the case of studies interested in chatbots for children and caregivers (Martin & Richmond, 2023).

Of the 53 studies included in Abd-Alrazaq et al. (2019), only 33 investigated effectiveness (of which 14 in a RCT). The lack of thorough validation in some chatbot solutions could hinder their immediate adoption for mental health. Furthermore, studies with small sample sizes represented in some instances a high proportion of the studies included in the reviews (e.g., 50 percent of studies had less than 50 participants in Abd-Alrazaq et al., 2019). This issue and the important proportion of pilot studies in most of the included reviews should be considered when assessing the effectiveness of chatbots.

The broader adoption of chatbots for mental health requires careful consideration of ethical issues. Several concerns have been highlighted in the literature, particularly regarding the evaluation of the platforms, including in terms of risks for the users and the question of information privacy, data protection and users' autonomy and consent (Fiske et al., 2019; Luxton & Hudlicka, 2022). The use of chatbots can lead to therapeutic misconceptions, i.e., users might misunderstand the scope and purpose of the chatbots if this information is not communicated clearly (Khawaja & Bélisle-Pipon, 2023). Additionally, the integration of chatbots into the healthcare system and the risk of users becoming overly reliant on such solutions remain significant points of debate (Fiske et al., 2019; Luxton & Hudlicka, 2022).

Strengths and limitations

The main strength of the current review of reviews on chatbots for mental health, published in scientific journals, lies in its ability to provide an overview of the field: We have summarized the scope and findings of existing reviews. A limitation of our review has been our focus on reviews published in English, as we might have missed reviews written in other languages because the search was performed in English only. Additionally, as most of the included reviews did not systematically focus on the effectiveness of chatbots, our results on this aspects stem mainly from the three included meta-analyses. Our review also did not concentrate on a specific population. The findings are thus not limited to specific groups, but are often reported in a general fashion which could reduce their specificity. Likewise, the absence of a requirement for the included reviews to focus only on studies including a comparator group may have only highlighted findings which could possibly also be present with other types of interventions. Further, owing to the paucity of studies (and hence reviews) comparing mental health chatbots to human-delivered psychotherapy (Bendig et al., 2022) or even internet-based therapy, we were unable to assess whether the effectiveness of chatbots is comparable to that of human therapists or internet interventions, which would be an important consideration for policymaking.

Conclusion

The majority of individuals presenting with mental disorder symptomatology do not obtain professional help (Henderson et al., 2013; WHO, 2021). This is partly due to a shortage of mental health professionals – leading to (anticipation of) long waiting periods, fear of stigma, distance, and lack of trust in mental health services (e.g., Cyr et al., 2019; Thornicroft, 2008; Williams et al., 2021). This review has summarized the scope and findings of existing reviews on chatbots for mental health published in academic journals. Studies with small sample sizes have been reported in most reviews. CBT has been found the intervention most chatbots rely upon. Depression and anxiety have been the most frequent targets of chatbots Correspondingly, the PHQ-9 and GAD-7 are the most commonly used measurement tools. The meta-analyses included in the review highlight the effectiveness of chatbots for depression and general distress, and potentially for other clinical target for which He et al.,

(2023) highlight the effectiveness of chatbots in studies with passive control groups but not active control groups.

While users generally have favorable perceptions of chatbots, some reviews have also indicated weaknesses, such as the lack of empathy and repetitiveness perceived by some users. We believe that advancements in these areas, as well as the inclusion of an assessment of users mental health and progress through interactions with the chatbots and a memory of such interactions (as in relational chatbots) necessary. The question of safety and data protection have also been raised. The aim of this scoping review was to describe the broad content of the current research on chatbots for mental health. Evaluating the quality of the reviews on this topic could be performed in a future systematic review.

Funding Open access funding provided by University of Basel. No funding has been obtained for this study.

Data availability The data extracted from the articles is provided within the manuscript in narrative/tabular forms.

Declarations

Ethical approval This scoping review of the literature does not require ethical approval as there are no participants.

Conflict of interests The authors declares no conflict of interest.

Informed consent This scoping review of the literature does not require an informed consent as there are no participants.

References

*Abd-Alrazaq, A. 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*, 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

*Abd-Alrazaq, A. 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),* e16021. https://doi.org/10.2196/16021

*Abd-Alrazaq, A. 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(5)*, e17828. https://doi.org/10.2196/17828

*Ahmed, A., Hassan, A., Aziz, S., Abd-alrazaq, A. A., Ali, N., Alzubaidi, M., Al-Thani, D., Elhusein, B., Siddig, M. A., Ahmed, M., & Househ, M. (2023). Chatbot features for anxiety and depression: A scoping review. *Health Informatics Journal, 29(1),* 14604582221146719.https://doi.org/10.1177/14604582221146719

Alonso, J., Liu, Z., Evans-Lacko, S., Sadikova, E., Sampson, N., Chatterji, S.,... WHO World Mental Health Survey Collaborators. (2018). Treatment gap for anxiety disorders is global: Results of the World Mental Health Surveys in 21 countries. *Depression and Anxiety, 35(3),* 195-208.https://doi.org/10.1002/da.22711

ASReview, LAB developers. (2022). ASReview LAB - A tool for AI-assisted systematic reviews. Zenodo. https://doi.org/10.5281/zenodo.7319063

*Bendig, E., Erb, B., Schulze-Thuesing, L., & Baumeister, H. (2022). The next generation: Chatbots in clinical psychology and psychotherapy to foster mental health – A scoping review. *Verhaltenstherapie, 32(2),* 64-76. https://doi.org/10.1159/000501812

Bickmore, T., & Gruber, A. (2010). Relational agents in clinical psychiatry. *Harvard Review of Psychiatry,18*(2), 119–130. https://doi.org/10.3109/10673221003707538

Bloom, S. M., White, R. J., Beckley, R. F., & Slack, W. V. (1978). Converse: A means to write, edit, administer, and summarize computer-based dialogue. *Computers and Biomedical Research,11*(2), 167–175. https://doi.org/10.1016/0010-4809(78)90028-9

Cole, R., Van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J.,... Yan, J. (2003). Perceptive animated interfaces: First steps toward a new paradigm for human-computer interaction. *Proceedings of the IEEE*, *91*(9), 1391–1405. https://doi.org/10.1109/JPROC.2003.817143

Cyr, M. E., Etchin, A. G., Guthrie, B. J., & Benneyan, J. C. (2019). Access to specialty healthcare in urban versus rural US populations: A systematic literature review. *BMC Health Services Research,19*(1), 1–17. https://doi.org/10.1186/s12913-019-4815-5

Darcy, A., Daniels, J., Salinger, D., Wicks, P., & Robinson, A. (2021). Evidence of human-level bonds established with a digital conversational agent: Cross-sectional, retrospective observational study. *JMIR Formative Research,5*(5), e27868. https://doi.org/10.2196/27868

Esfandiari, N., Mazaheri, M. A., Akbari-Zardkhaneh, S., Sadeghi-Firoozabadi, V., & Cheraghi, M. (2021). Internet-delivered versus face-to-face cognitive behavior therapy for anxiety disorders: systematic review and meta-analysis. *International Journal of Preventive Medicine*, *12*. https://doi.org/10.4103/ijpvm.ijpvm_208_21

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research,21*(5), e13216. https://doi.org/10.2196/13216

*Gaffney, H., Mansell, W., & Tai, S. (2019). Conversational agents in the treatment of mental health problems: Mixed-method systematic review. *JMIR Mental Health, 6(7),* e14166. https://doi.org/10.2196/14166

GBD 2019 Mental Disorders Collaborators. (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry*, *9*(2), 137–150. https://doi.org/10.1016/S2215-0366(21)00395-3

Grodniewicz, J. P., & Hohol, M. (2024). Therapeutic chatbots as cognitive-affective artifacts. *Topoi,43*, 795–807. https://doi.org/10.1007/s11245-024-10018-x

Grolleman, J., van Dijk, B., Nijholt, A., van Emst, A. (2006). Break the Habit! Designing an e-Therapy Intervention Using a Virtual Coach in Aid of Smoking Cessation. In: IJsselsteijn, W. A., de Kort, Y. A. W., Midden, C., Eggen, B., van den Hoven, E. (Eds.) *Persuasive technology. PERSUASIVE 2006. Lecture notes in computer science,* vol 3962. Springer. https://doi.org/10.1007/11755494_19

*He, Y., Yang, L., Qian, C., Li, T., Su, Z., Zhang, Q., & Hou, X. (2023). Conversational agent interventions for mental health problems: Systematic review and meta-analysis of randomized controlled trials. *Journal of Medical Internet Research, 25(9),* e43862. https://doi.org/10.2196/43862

Henderson, C., Evans-Lacko, S., & Thornicroft, G. (2013). Mental illness stigma, help seeking, and public health programs. *American Journal of Public Health,103*(5), 777–780. https://doi.org/10.2105/AJPH.2012.301056

Herbener, A. B., Klincewicz, M., & Damholdt, M. F. (2024). A narrative review of the active ingredients in psychotherapy delivered by conversational agents. *Computers in Human Behavior Reports*, 100401. https://doi.org/10.1016/j.chbr.2024.100401

Hussain, S., Ameri Sianaki, O., Ababneh, N. (2019). A survey on conversational agents/chatbots classification and design techniques. In: Barolli, L., Takizawa, M., Xhafa, F., Enokido, T. (Eds.), Web, *Artificial Intelligence and Network Applications*. WAINA 2019. Advances in intelligent systems and computing (Vol. 927). Springer. https://doi.org/10.1007/978-3-030-15035-8_93

*Jabir, A. I., Martinengo, L., Lin, X., Torous, J., Subramaniam, M., & Car, L. T. (2023). Evaluating conversational agents for mental health: Scoping review of outcomes and outcome measurement instruments. *Journal of Medical Internet Research, 25(4),* e44548. https://doi.org/10.2196/44548

Karyotaki, E., Efthimiou, O., Miguel, C., genannt Bermpohl, F. M., Furukawa, T. A., Cuijpers, P.,... Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration. (2021). Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. *JAMA Psychiatry*, *78*(4), 361–371. https://doi.org/10.1001/jamapsychiatry.2020.4364

Khawaja, Z., & Bélisle-Pipon, J. C. (2023). Your robot therapist is not your therapist: Understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health,5*, 1278186. https://doi.org/10.3389/fdgth.2023.1278186

Kocaballi, A. B., Berkovsky, S., Quiroz, J. C., Laranjo, L., Tong, H. L., Rezazadegan, D.,... Coiera, E. (2019). The personalization of conversational agents in health care: systematic review. *Journal of Medical Internet Research*, *21*(11), e15360. https://doi.org/10.2196/15360

Konstantinidis, E. I., Hitoglou-Antoniadou, M., Luneski, A., Bamidis, P. D., & Nikolaidou, M. M. (2009). Using affective avatars and rich multimedia content for education of children with autism. In *Proceedings of the 2nd international conference on pervasive technologies related to assistive environments* (pp. 1–6). ACM. https://doi.org/10.1145/1579114.1579172

Ku, J., Han, K., Lee, H. R., Jang, H. J., Kim, K. U., Park, S. H.,... Kim, S. I. (2007). VR-based conversation training program for patients with schizophrenia: a preliminary clinical trial. *Cyberpsychology & Behavior*, *10*(4), 567–574. https://doi.org/10.1089/cpb.2007.9989

Lim, S. M., Shiau, C. W. C., Cheng, L. J., & 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

Luxton, D. D., & Hudlicka, E. (2022). Intelligent virtual agents in behavioral and mental healthcare: Ethics and application considerations. In *Artificial intelligence in brain and mental health: Philosophical, ethical & policy issues* (pp. 41–55). Springer International Publishing

*Martin, R., & Richmond, S. (2023). Conversational agents for Children's mental health and mental disorders: A scoping review. *Computers in Human Behavior: Artificial Humans, 1,* 100028.https://doi.org/10.2196/44548

*Martínez-Miranda, J. (2017). Embodied conversational agents for the detection and prevention of suicidal behaviour: Current applications and open challenges. *Journal of Medical Systems, 41(7),* 135.https://doi.org/10.1007/s10916-017-0784-6

Moore, J. R., & Caudill, R. (2019). The bot will see you now: A history and review of interactive computerized mental health programs. *Psychiatric Clinics,42*(4), 627–634. https://doi.org/10.1016/j.psc.2019.08.007

*Ogilvie, L., Prescott, J., & Carson, J. (2022). The use of chatbots as supportive agents for people seeking help with substance use disorder: A systematic review. *European Addiction Research, 28(6),* 405-418. https://doi.org/10.1159/000525959

Osgood-Hynes, D. J., Greist, J. H., Marks, I. M., Baer, L., Heneman, S. W., Wenzel, K. W., Manzo, P. A., Parkin, J. R., Spierings, C. J., Dotti, S. L., & Vitse, H. M. (1998). Self-administered psychotherapy for depression using a telephone-accessed computer system plus booklets: An open US-UK study. *Journal of Clinical Psychiatry,59*(7), 358–365. https://doi.org/10.4088/jcp.v59n0704

*Otero-González, I., Pacheco-Lorenzo, M.R., Fernández-Iglesias, M.J., & Anido-Rifón, L.E. (2024). Conversational agents for depression screening: a systematic review. *International Journal of Medical Informatics, 181*, 105272. https://doi.org/10.1016/j.ijmedinf.2023.105272

*Pacheco-Lorenzo, M. R., Valladares-Rodríguez, S. M., Anido-Rifón, L. E., & Fernández-Iglesias, M. J. (2021). Smart conversational agents for the detection of neuropsychiatric disorders: a systematic review. *Journal of Biomedical Informatics*, *113*, 103632. https://doi.org/10.1016/j.jbi.2020.103632

*Provoost, S., Lau, H. M., Ruwaard, J., & Riper, H. (2017). Embodied conversational agents in clinical psychology: A scoping review. *Journal of Medical Internet Research, 19(5),* e151. https://doi.org/10.2196/jmir.6553

Seeger, A. M., Pfeiffer, J., & Heinzl, A. (2021). Texting with humanlike conversational agents: Designing for anthropomorphism. *Journal of the Association for Information Systems,22*(4), 8. https://doi.org/10.17705/1jais.00685

Selmi, P. M., Klein, M. H., Greist, J. H., Sorrell, S. P., & Erdman, H. P. (1990). Computer-administered cognitive-behavioral therapy for depression. *The American Journal of Psychiatry,147*(1), 51–56. https://doi.org/10.1176/ajp.147.1.51

Slack, W. V. (2000). Patient-computer dialogue: A review. *Yearbook of Medical Informatics,9*(01), 71–78. https://doi.org/10.1055/s-0038-1637944

Slack, W. V., & Slack, C. W. (1977). Talking to a computer about emotional problems: A comparative study. *Psychotherapy: Theory, Research & Practice,14*(2), 156–164. https://doi.org/10.1037/h0086523

Thornicroft, G. (2008). Stigma and discrimination limit access to mental health care. *Epidemiology and Psychiatric Sciences,17*(1), 14–19. https://doi.org/10.1017/S1121189X00002621

*Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *The Canadian Journal of Psychiatry, 64(8),* 456-464.https://doi.org/10.1177/0706743719828977

*Vaidyam, A. N., Linggonegoro, D., & Torous, J. (2020). Changes to the psychiatric chatbot landscape: A systematic review of conversational agents in serious mental illness. *The Canadian Journal of Psychiatry, 66(5),* 339-348.https://doi.org/10.1177/0706743720966429

Wang, X., Luo, R., Liu, Y., Chen, P., Tao, Y., & He, Y. (2023). Revealing the complexity of users' intention to adopt healthcare chatbots: A mixed-method analysis of antecedent condition configurations. *Information Processing & Management,60*(5), 103444. https://doi.org/10.1016/j.ipm.2023.103444

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM,9*(1), 36–45. https://doi.org/10.1145/365153.365168

WHO. (2021). *World mental health 2020*. World Health Organization.

Williams, L., Hayes, G., Washington, G., W. Black, R., Williams, C., Clements, L., & Allotey, M. (2021). Analysis of distance-based mental health support for underrepresented university students. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–6). https://doi.org/10.1145/3411763.3451708

Zarr, M. L. (1984). Computer-mediated psychotherapy: Toward patient-selection guidelines. *American Journal of Psychotherapy,38*(1), 47–62. https://doi.org/10.1176/appi.psychotherapy.1984.38.1.47

Zhang, W., Yang, W., Ruan, H., Gao, J., & Wang, Z. (2023). Comparison of internet-based and face-to-face cognitive behavioral therapy for obsessive-compulsive disorder: A systematic review and network meta-analysis. *Journal of Psychiatric Research,168*, 140–148. https://doi.org/10.1016/j.jpsychires.2023.10.025

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.