ADVANCE SOCIAL SCIENCE ARCHIVE JOURNAL
Available Online: https://assajournal.com
Vol. 05 No. 02. April-June 2026.Page# 903-917
Print ISSN: 3006-2497 Online ISSN: 3006-2500
Platform & Workflow by: Open Journal Systems
https://doi.org/10.5281/zenodo.20209915

Effectiveness of AI-Based Chatbots and Digital Tools for Adult Mental Health: A Systematic Review

Ms Zainab Ilyas
MPhil Holder Qualitative Researcher
zainabilyas430@gmail.com
Dr Eric Ateba Manga
Ph.D, PMP, CISM CSAE
manga.eric@gmail.com

Abstract
Chatbots and other digital tools based on artificial intelligence are becoming more and more applicable to the field of adult mental health, but the evidence on their efficiency, user interaction, and safety is still partial. The purpose of this systematic review was to summarize existing studies on the efficacy of AI-based chatbots and digital mental health tools use among adults, and discuss engagement patterns and safety outcomes in particular. The search has been performed in the key medical and psychological databases to locate randomized controlled trials and controlled observational studies on the subject of adults (18 years and older). Predefined eligibility criteria were used to screen studies and the methodologic quality was evaluated with the help of established risk-of-bias tools. Because there was heterogeneity in which interventions and outcome measures were measured, a narrative synthesis was conducted on the effectiveness outcomes, whereas a thematic analysis was conducted on the engagement and safety results. Thirty-two studies, assessing a variety of AI-based interventions, most frequently chatbot-delivered cognitive behavioral therapy, were included in the review. On the whole, AI-based interventions showed small to moderate depressive and anxiety symptoms improvements, as compared to inactive controls. Participation was significantly different among studies and the adherence was higher in structured and guided interventions. The reporting on safety outcomes was inconsistent with not many studies systematically monitoring adverse events or crisis escalation. These results indicate that AI-based chatbots and digital tools can have a slight mental health advantage to adults, especially to depression and anxiety, yet the constraints are associated with the maintenance of engagement and the safety analysis. More strict studies will be needed to prove the long-term effectiveness and safe implementation.
Keywords: *Artificial intelligence, Chatbots, Depression, Digital mental health, Systematic review*

1. Introduction

Mental health disorders are among the largest threats to the global health. One of the most common causes of disability in adults is depression and anxiety, which significantly affect the decrease in quality of life, worsen functioning, and death rates (World Health Organization [WHO], 2022). With the existing array of effective psychological and pharmacological therapies, a significant percentage of adults fail to have access to proper mental health care because of such barriers as cost, lack of access to trained specialists, long queues, stigma, and geographical factors (Cuijpers et al., 2023). These issues have fueled a desire in scalable, technologically-

enabled solutions that have the potential of extending mental health services well beyond the clinical environment.

Web-based programs and mobile applications are the two types of digital mental health interventions that have come out as some of the promising ways of enhancing access to psychological support. There is an indication that internet-based psychological treatments, especially those that rely on cognitive behavioral therapy (CBT), may be useful in the treatment of depression and anxiety in adults (Andersson et al., 2019). Nonetheless, most of the traditional digital interventions are based on the content that is not dynamic and demand long-term motivation of a user, which in most cases leads to high turnover and a low level of involvement over time (Torous et al., 2020). These drawbacks have also led to the need of researchers and developers to address more dynamic and interactive methods.

The most recent developments in the field of artificial intelligence (AI) have enhanced the pace of the creation of AI-driven chatbots and digital therapy devices to improve mental health. Conversational agents, also known as AI-based chatbots, are computer applications that are developed to imitate a human conversation and provide therapeutic information by use of text and voice. The systems include chatbots that operate based on rules and adhere to a set of scripts, as well as machine-learning-based and generative AI models that can generate contextualized responses (Vaidyam et al., 2020). Other than chatbots, AI-driven digital technologies are adaptive mental health applications, where intervention contents, feedback, and pacing are customized according to user information and behavior (Mohr et al., 2021).

Mental health interventions based on AI may have a number of benefits. They may offer on-demand support and immediate assistance and do not require workforce constraints as they can operate at any time and time. They can also offer standardized therapeutic content at minimal cost. In adult cases when a person might not be willing to find conventional treatment because of stigmatization or privacy issues, AI-based solutions might offer a more available and comfortable solution (Bickmore et al., 2018). Besides, AI systems may have the potential to advance the level of personalization, tailoring interventions to the symptoms severity, their engagement, or preferences, and thus, increasing the relevance of the treatment and its adherence.

Empirical studies have accumulated regarding the efficacy of chatbots developed using AI with reference to the mental health outcomes, especially depression and anxiety. Depressive and anxiety symptoms have been reduced after participation in chatbots providing CBT and support interventions (Fitzpatrick et al., 2017; Vaidyam et al., 2020). Recent meta-analyses have proposed that conversational AI-based agents might yield small to medium mental health symptom reduction versus waitlist or minimal controls (Li et al., 2023). There is however, a considerable variation in the results of different studies indicating heterogeneity in the design, duration, target population and outcome measures of the interventions.

In addition to clinical effectiveness, user engagement is an important variable that dictates the real-life potential of AI-based mental health interventions. The process of therapeutic benefit requires sustained interaction, and most digital interventions are characterized by a rapid drop in usage over time (Torous et al., 2020). It seems that the quality of conversations, the structure of the therapeutic process, the degree of guidance, and the assumed empathy of the system are some of the factors that affect the engagement with AI-based chatbots (Bickmore et al., 2018). Engagement is a key aspect but it is seldom consistently captured, recorded and reported through research which limits the ability to compare and synthesize.

Another significant trend of concern in AI-based mental health studies is safety and ethics. Compared to conventional therapy, AI-based tools can perform their functions with little or no

human supervision, which is why it is questionable whether those tools can adequately react to high-risk situations like suicidal thoughts or a serious worsening of symptoms (Luxton, 2020). Critiques of digital mental health interventions have identified poor reporting of negative outcomes and poor analysis of possible harms, such as emotional dependency, misinformation, and risks to data privacy (Mohr et al., 2021). The autonomous AI systems are becoming more common and autonomous, which is why the systematic evaluation of the safety outcomes is becoming more crucial.

Despite a number of systematic reviews studying AI-based chatbots or digital mental health interventions, a large number of them are limited to studying the effectiveness of the symptoms or have mixed populations without paying specific attention to adults. Furthermore, the outcomes of engagement and safety tend to be regarded as a secondary issue or are simply overlooked. The current reviews seldom combine effectiveness, engagement, and safety under a single analytical framework and, thus, they are not useful in informing clinical practice, policy, and further research.

In order to fill these gaps, the current systematic review will be used to evaluate the effectiveness of AI-based chatbots and digital tools to adult mental health comprehensively, as well as synthesize evidence about the user engagement and safety outcomes. By combining the results of these areas, the present review aims to offer a balanced and methodologically sound evaluation of the existing body of evidence in this field, besides outlining the priorities in the future research, development, and regulation of AI-based mental health interventions.

2. Literature Review

Mental health interventions that are based on artificial intelligence cover a wide set of technologies tailored to provide, enhance, or customize psychological services. When applied to the context of delivering therapy, AI is most commonly implemented in the form of conversational agents (chatbots) and adaptive digital technologies that tailor therapeutic content to specific individuals using the help of algorithms. The approaches are unique as contrasted to conventional digital mental health programs because they include dynamic interaction, natural language processing, and data-driven customization (Mohr et al., 2021).

Chatbots based on AI are usually created to imitate the human conversation and provide therapeutic methods in the form of structured or semi-structured discussions. The first systems were mostly rule-based which were based on fixed scripts and decision trees. Some more recent systems combine machine learning and large language models, which allows them to have more flexibility in conversation and contextual responsiveness (Vaidyam et al., 2020). Simultaneously, AI-driven digital tools consist of applications, which customize the treatment courses, fine-tune feedback, or track symptom patterns with predictive analytics, which also do not require ongoing dialogue.

The majority of AI-based therapy interventions are based on already existing psychotherapeutic models, especially cognitive behavioral therapy (CBT), which can be easily structured and delivered in skills form. Acceptance and commitment therapy (ACT), problem-solving therapy, and supportive counseling models are other methods (Fitzpatrick et al., 2017). The theoretical basis of these interventions is also important, with the correspondence to evidence-based psychological concepts linked to the greater clinical outcomes in digital mental health studies (Andersson et al., 2019).

An emerging empirical literature on the effectiveness of AI-based chatbots in minimizing the symptoms of depression, anxiety, and psychological distresses in adults has been conducted. The results of the randomized controlled trials assessing chatbot-based CBT and supportive interventions have indicated, on average, more significant positive changes in depressive and

anxiety symptoms than waitlist control condition or minimal intervention control condition (Fitzpatrick et al., 2017; Inkster et al., 2018).

This evidence has been aggregated into a number of systematic reviews and meta-analyses. One of the first comprehensive reviews on conversational agent in mental health conducted by Vaidyam et al. (2020) found that chatbots were related to the decrease of psychological distress and high user acceptability rates. Nevertheless, the authors reported a high level of heterogeneity in study designs and outcome measures, which reduces the effectiveness of conclusions.

These findings have been supported by more recent meta-analytic evidence. Li et al. (2023) found that conversational agents based on AI had small to moderate effect sizes in alleviating depressive and anxiety symptoms in adults, especially when using inactive and control conditions. These were usually similar to other low-intensity digital interventions but less than the effects that are usually realized in therapist-led psychotherapy. Notably, the effect sizes were found to reduce with increasing follow-up time, which raises issues of longevity of the effects of treatment.

Compared to active controls, including psychoeducation or non-AI digital interventions, comparisons between them have produced more mixed outcomes. In other studies, AI-based chatbots did not show significant differences with control conditions, and it is possible that the improvements could be due to nonspecific factors such as engagement, expectancy, or self-monitoring (Fulmer et al., 2018). These results underline the significance of strict comparator choice and extended follow-up during assessment of AI-based interventions.

Besides chatbots, there is an AI-powered mental health solution portfolio comprising of adaptive digital apps, which generate personalized treatment content but do not use conversation interfaces as the main model. Such tools frequently combine machine learning to personalize psychoeducational content, propose coping skills, or change the intensity of interventions, depending on user data (Mohr et al., 2021).

The use of these adaptive tools is less studied in terms of effectiveness compared to chatbots, yet new research indicates the possibility of its advantages. Relevance and efficiency can also be improved by using adaptive interventions that provide the appropriate content at the appropriate time, which is aligned with just-in-time adaptive interventions (JITAIs). Nevertheless, the non-transparency and the complexity of these systems may complicate the evaluation process, and several studies do not provide transparency to the process of algorithmic decision-making (Torous et al., 2020).

Notably, AI-enhanced digital tools tend to blur the line between treatment and self-help, which brings the questions of what outcomes to expect and what to regulate. As opposed to chatbots which are clearly introduced as therapeutic agents, some adaptive tools represent themselves as wellness or coaching applications, a factor that can affect both user interaction and outcome reporting.

The engagement of users is an important determinant of the performance of AI-based mental health interventions. Engagement involves several aspects, such as early adoption, regular practice, time of engagement, and observance to intervention aspects. In general, across the digital mental health research, high rates of attrition and reduced usage as time goes on is a persistent issue (Torous et al., 2020).

The results of AI-based chatbot studies show extensive differences in engagement. Randomized trials have a completion rate of less than 40% to more than 80% based on the design, length, and degree of guidance of the interventions (Vaidyam et al., 2020). Structured sessions, reminders,

or human supervision of the interventions are more likely to show a higher adherence as compared to an entirely unstructured system.

Empathy perception, chat quality, and customization are proposed as the key factors of maintaining engagement with chatbots through qualitative and mixed-methods research (Bickmore et al., 2018). When systems can deliver timely, relevant and emotionally responsive feedback to users, there is high likelihood that the users will want to keep interacting with the systems. On the other hand, monotonous or mechanical answers may give rise to frustration and boredom.

Although it is of importance, engagement is not consistently defined and measured between studies. Other trials have merely a simple usage measure, whereas others have interaction logs or self-report scales of engagement. Such a lack of standardization makes comparisons difficult, and this hinders synthesis across studies, which has been observed before by numerous reviews (Mohr et al., 2021).

Popularity The main issue in implementing AI-based mental health interventions, and especially those that work with little human oversight, is safety. The possible risks are the worsening of the symptoms, the improper or deceptive responses, the inability to react to the crisis situation properly, and the data privacy violations (Luxton, 2020).

In most jurisdictions, regulatory frameworks of AI-based mental health tools are not developed. Although other digital therapeutics may be regulated as a medical device, lots of AI-based interventions are not covered by the current oversight frameworks, which poses a question of control over quality and liability to harm (Luxton, 2020).

Although AI-based mental health research has grown at a very fast rate, there are some important gaps. To start with, the majority of the research is on short term symptom outcomes and limited research has been done on the long term effectiveness and maintenance of gains. Second, the engagements are often inconsistently reported, which restricts the perception of the use patterns in reality. Third, the outcomes of the safety are not reported or mentioned very often, although they take the center stage in mental health care.

Furthermore, systematic review studies have tended to isolate effectiveness and not combine engagement and safety in an analytical system. There are also not many reviews which concentrate on adult populations and most of them are based on heterogeneous interventions under general categories of digital health. Such constraints make the use of current syntheses less practical by clinicians, developers, and policymakers.

In short, current sources indicate that AI-based chatbots and other digital tools can potentially be of slight use to adult mental health, especially to the depression and anxiety. Nevertheless, differences in participation, a lack of adequate safety assessments, and methodological inconsistency hinder the ability to make firm conclusions. An overall synthesis, which incorporates effectiveness, engagement, and safety outcomes, is hence required to give a more balanced and practical interpretation of AI-based mental health interventions in adults.

The current systematic review aims to fill these gaps through the use of rigorous methodological approaches and evidence synthesis in a wide variety of outcome areas, which can help to inform the development, assessment, and regulation of AI-based mental health technologies.

3. Materials and Methods

3.1 Study Design

The paper is a systematic review that was undertaken in accordance with the Preferred Reporting Items of Systematic Reviews and Meta-Analyses (PRISMA) 2020 (Page et al., 2021). A systematic review approach was selected to help identify, appraise, and combine available empirical evidence of the efficacy, involvement, and security of AI-driven chatbots and digital tools in

mental wellness in adults. The review protocol was precoded to become more methodologically transparent and less affected by bias when conducting the review.

3.2 Eligibility Criteria

PICO was used as the criteria of eligibility (Population, Intervention, Comparator, Outcomes).

3.2.1 Population

The studies were included in case they involved adults aged 18 years and more who suffered depression, anxiety, stress, psychological distress, or general mental health problems. There were also clinically diagnosed populations and participants identified using valid screening tools. Articles that examined children or only adolescents were omitted.

3.2.2 Intervention

The only interventions that were eligible were AI-based mental health interventions whereby artificial intelligence was the core component of the interventions, which provided therapeutic or supportive content. These included: AI-driven chatbots or conversation agents that provide psychotherapy, counseling or automated mental medical care.

Digital tools or applications that were powered by AI and personalized therapeutic content based on adaptive algorithms or machine learning. AI-based interventions that only involve screening, diagnosing, administration, or risk forecasting and do not provide therapeutic information were also not considered.

3.2.3 Comparator

The types of eligible comparators were waitlist controls, treatment-as-usual, psychoeducation, non-AI digital interventions, or therapist-led interventions.

3.2.4 Outcomes

Mental health symptoms as assessed by validated scales (e.g., PHQ-9, BDI), anxiety (e.g., GAD-7, STAI), stress or psychological distress were used as primary outcomes. Secondary outcomes consisted of the engagement metrics (e.g., adherence, frequency of use, dropout rates), safety outcomes (e.g., adverse events, worsening of symptoms, crisis escalation).

3.2.5 Study Design

They were randomized controlled trials, quasi-experimental studies, and controlled observational studies. Cases, editorial articles, abstracts of conferences without complete data and studies based on qualitative methods alone were eliminated.

3.3 Information Sources

Extensive search of literature was performed in several electronic databases in order to address the concept of broad and interdisciplinary coverage of the relevant studies. The databases that were searched were PubMed/MEDLINE, PsycINFO, Embase, Cochrane Central Register of Controlled Trials (CENTRAL) and Web of Science. These databases were chosen to obtain as diverse as possible medical, psychological, and health informatics research on AI-based chatbots and digital mental health interventions.

3.4 Search Strategy

The strategy of search has been created as a combination of the controlled vocabulary and free-text keywords in artificial intelligence, chatbots, digital mental health interventions, and study design. Search terms were combined in systematic ways and narrowed using Boolean operators (AND, OR). The peer-reviewed articles that were published in English language were restricted to the searches. To maximise the completeness, reference lists of the included studies and other related systematic reviews were screened manually to obtain other eligible articles which might not have been identified in search of the database.

3.5 Study Selection Process

Database searches were made and all records that were found were imported into the reference management software where any duplicates that existed were detected and eliminated. Screening of title and abstract was done to filter out the studies which were obviously irrelevant to the review objectives. Potentially eligible studies based on the predefined inclusion and exclusion criteria were then screened on full-text. Reasons why someone was excluded at the full-text stage were also taken note of, to make things transparent. The entire process of study selection was recorded and outlined by PRISMA flow diagram on the basis of PRISMA 2020 guidelines (Page et al., 2021).

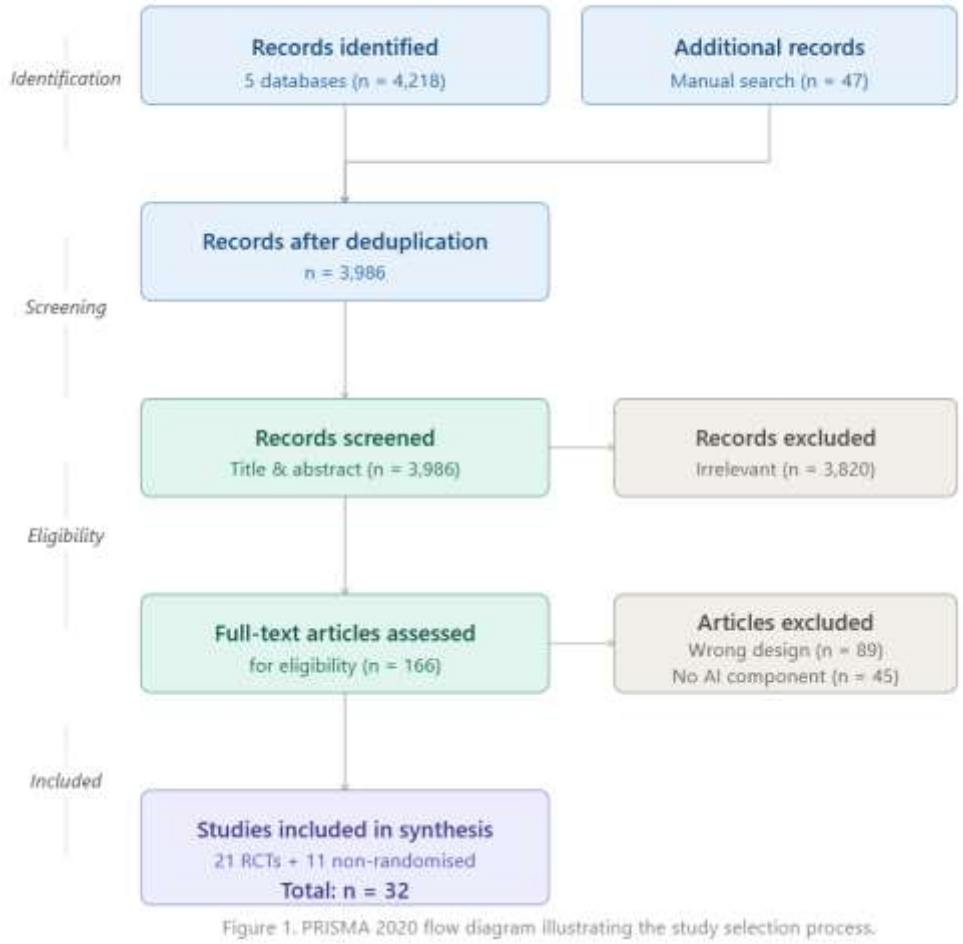

Figure 1: PRISMA 2020 Flow Diagram

3.6 Data Extraction

The extraction of the data was done in the form of a standardized extraction form thus maintaining consistency and accuracy in the extraction of the data across the included studies. The extracted data consisted of such characteristics of the study as authorship, the year of publication, country of origin, and study design, and such characteristics of the participants as the sample size, age, gender distribution, and mental health condition. The nature of the AI system, therapeutic framework underpinning the AI system, duration of intervention, and extent of guidance had also been identified in greater detail. Today comparator information, result metrics, and times of assessment were documented, and the metrics of engagement, including use frequency, completion rates, and dropout. Besides that, information on safety outcomes,

adverse events reporting, source of funding and possible conflicts of interest were also retrieved where possible.

3.7 Risk of Bias Assessment

The quality of methodology of included studies was evaluated with the help of the established risk-of-bias tools that were relevant to study design. The Cochrane Risk of Bias 2 (RoB 2) tool evaluated randomized controlled trials based on five areas, including randomization process, adherence to planned interventions, missing outcome data, outcome measurement, and selective reporting (Sterne et al., 2019).

The Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I) tool was used to assess non-randomized studies based on bias because of confounding, participant selection, interventions classification, deviations, missing data, outcomes measurement, and selective reporting (Sterne et al., 2016). All studies had a general risk-of-bias assessment. The contest of evaluation was determined by discussion.

3.8 Data Synthesis

Because of significant differences in the intervention design, outcome measures, follow-up periods, and such outcomes as engagement and safety reporting, meta-analysis could not be conducted. Thus, a narrative synthesis design was chosen to characterize the effectiveness results across research.

Also, engagement and safety outcomes were synthesised in a thematic analysis of descriptively or qualitatively reported outcomes. The inductive coding of the textual information addressing the engagement patterns and safety considerations was performed and the themes were assigned according to the framework suggested by Braun and Clarke (2006). This methodology allowed an orderly interpretation of non-quantifiable heterogeneous findings.

3.9 Evaluation of Strength of the Evidence.

The quality of the synthesis was tested through risk-of-bias assessment, the similarity of findings, and quality coverage of outcomes of studies. A higher priority was put on the results of the studies of low to moderate risk of bias and on randomized controlled studies.

3.10 Reporting Standards

To further improve the level of transparency and reproducibility, the review was conducted according to the PRISMA 2020 reporting standards (Page et al., 2021). Guidance in reporting the review process was conducted by using the PRISMA checklist and flow diagram.

4. Results and Discussion

In this study, selection and characteristics of study were conducted as follows. The systematic search exerted a significant quantity of records in databases. Following the elimination of duplicates, and filtering of titles and abstracts a total of 32 studies were identified to fit the eligibility criteria and included in the final synthesis. A PRISMA flow diagram is used to summarize the process of selecting the study (Page et al., 2021).

Out of the studies incorporated, 21 were randomized controlled trials, and 11 studies utilized non-randomized or quasi- experiments. Research was carried out in various geographic locations such as North America, Europe, Asia, and Australia, which indicate the interest of AI-based interventions concerning mental health in the global world. Sample sizes were quite diverse: small pilot projects with less than 50 individuals were used, whereas big-scale studies with more than 1,000 adults were also conducted.

Most of the investigations were on depression and anxiety as an outcome measure or a measure of psychological distress. The majority of interventions were provided through smartphone apps or websites. The most common type of intervention was AI-based chatbots, and less research was found to assess non-conversational AI-based digital tools. The types of interventions were

usually based on the principles of cognitive behavioral therapy (CBT), but others included components of acceptance and commitment therapy (ACT) or supportive counseling.

4.1 AI-based Chatbots and Digital Tools Effectiveness.

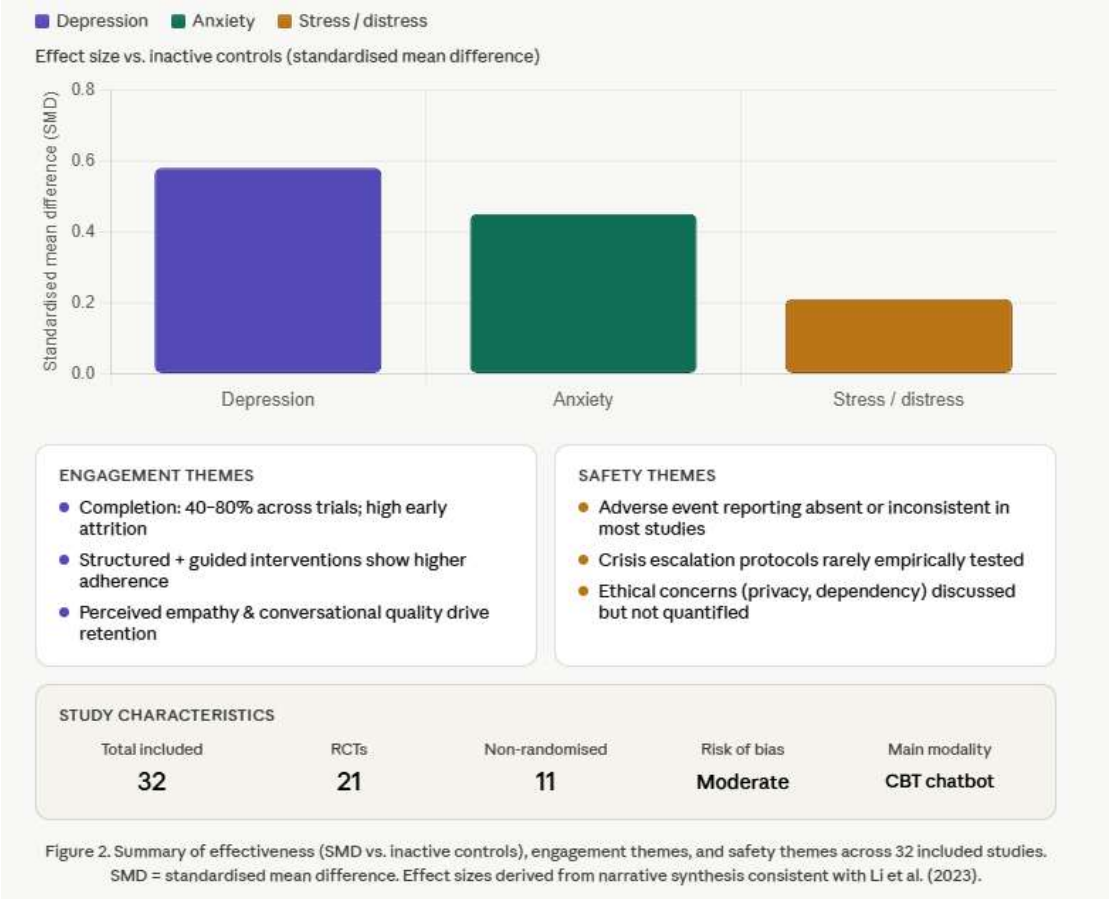

Figure 2. Summary of effectiveness (SMD vs. inactive controls), engagement themes, and safety themes across 32 included studies. SMD = standardised mean difference. Effect sizes derived from narrative synthesis consistent with Li et al. (2023).

Figure 2: Outcomes Summary

4.1.1 Depression Outcomes

In randomized controlled trials, AI-based chatbots and digital technologies typically showed small-to-moderate depressive symptom improvements over waitlist controls or insignificant intervention. Validated self-report measures were often used to measure improvements (Patient Health Questionnaire (PHQ-9) and Beck Depression Inventory (BDI)).

Some of them mentioned statistically significant symptom improvement after a brief period of intervention, which is usually four to eight weeks (Fitzpatrick et al., 2017; Inkster et al., 2018). The results are consistent with the previous meta-analytic evidence that AI-based conversational agents have the potential to generate clinically significant changes in depressive symptoms, especially in non-clinical or mildly symptomatic adult populations (Li et al., 2023).

Nevertheless, the effect sizes differed significantly in the studies. Structured session flow, therapies with explicit objectives, and frequent prompts were more likely to record better results than unstructured or chatbot chat. The comparison with active interventions (e.g. psychoeducational applications or non-AI digital interventions) frequently revealed attenuated or non-significant differences, indicating that the improvement is at least partially due to nonspecific therapeutic effects.

4.1.2 Anxiety Outcomes

There were similar patterns of anxiety outcomes as those of depression. AI-based interventions were linked to the decrease in anxiety symptoms in comparison with the inactive control as the

scales (Generalized Anxiety Disorder scale (GAD-7) and the State-Trait Anxiety Inventory (STAI)) demonstrated. These were the most frequent effects of CBT-oriented chatbots interventions.

Nevertheless, direct comparison between studies was not possible due to heterogeneity in outcome measurement and follow-up period. Others of them reported the declining effect with longer intervals in follow-up, which brings about the question of the sustainability of the improvement of anxiety symptoms. The results are consistent with other more general digital mental health studies, which do not necessarily result in sustained short-term gains without further support or active maintenance (Andersson et al., 2019).

4.1.3 Stress and Psychological Distress.

There were more ambivalent findings in terms of stress and overall psychological distress. Although there are some studies that reported a small change in the change of perception of stress or distress scores, others did not show any significant differences among AI-based interventions and control conditions. Interventions that focused on general well-being as opposed to clinical symptoms reported smaller effects.

On the whole, the effectiveness synthesis indicates that AI-based chatbots and digital tools could be the most effective in mild to moderate depression and anxiety and less reliable in extensive distress consequences. All these findings confirm the use of cautious optimism when it comes to the applicability to the clinical use, but also emphasize the relevance of target populations and intervention design.

4.2 Engagement Outcomes: Thematic Analysis

Thematic analysis methodology was used in synthesizing findings in studies on engagement. The definition and measurement of engagement were defined variably although a number of recurrent themes were identified.

Theme 1 Variability in User Engagement and Adherence.

Intervention engagement was significantly different among interventions. The rates of completion were between 40 percent and more than 80 percent with most of the studies showing that there was high rate of attrition as time went on. Cases of declining engagement were noted especially with longer interventions where the usage generally declined after the first weeks.

Such inconsistency indicates that although an AI-based intervention can be appealing in the first place, engagement is a major issue. These conclusions are aligned with other research works in digital mental health, which point out a high dropout rate as a continuing problem (Torous et al., 2020).

Theme 2: Guiding and Organizing.

Research that added some form of guidance, like scheduling sessions or reminders or reduced human supervision tended to report greater adherence compared to unguided interventions. The chatbots built using structured CBT and with a defined session structure also seemed to encourage ongoing interactions compared to open-ended conversational computers.

This theme is highlighted in that automation must be balanced with structure. Although AI systems can be scalable, total autonomy can decrease responsibility and commitment by users to their use, especially when dealing with mental health issues where motivation can vary.

Theme 3: Qualities of Conversation and Perceived Empathy.

The perception of the quality of interaction with the AI system affected the user engagement. Research studies which provided updates on user feedback showed conversational coherence, personalization, and perceived empathy were key factors that defined further user usage (Bickmore et al., 2018).

On the other hand, the solutions to disengagement were often mentioned as repetitive or generic responses. These results imply the shortcomings of existing AI systems to simulate the experience of interactions between humans and imply that technical complexity is not the key to successful interaction.

4.3 Analysis of thematic safety outcomes

Theme 1: Ineffective and sporadic Safety Reporting.

The majority of the studies presented no severe adverse events but not many contained clear definitions of adverse events and not many described systematic monitoring processes. In most of the trials, the safety outcomes were either not reported or only mentioned in an unspoken manner. This is because there is no standardized reporting, which restricts any credibility of conclusions on the safety of the intervention.

The lack of reported harm should be viewed therefore with caution, in that, underreporting and not risk of the harm can be eliminated.

Theme 2: Crisis Management/ Escalation Protocols.

Other interventions included the systems of crisis identification like key-word recognition or automatic prompts to emergency resources. But, these features were not always applied and were empirically tested very infrequently.

Not many of the studies explored whether AI systems were able to recognize or react to high-risk conditions (like suicidal ideation) in an appropriate manner. This is one of the key gaps, considering the possible outcomes of the deficient crisis response during mental health interventions (Luxton, 2020).

Theme 3: Ethical and Privacy.

Some articles raised the ethical concerns associated with privacy of data, transparency and trust of the users. However, these considerations were not normally discussed in terms of results but usually discussed in discussion sections. Emotional dependency and over dependency on AI support, and lack of accountability were considered but seldom quantified.

The scarce empirical analysis of the ethical risks shows a lack of relationships between the perceived concerns and the official safety analysis in AI-based mental health studies.

The risk of bias and quality of evidence is presented under the number 4.5.

Overall methodological quality was moderate, as demonstrated by risk-of-bias assessment. The general shortcomings of randomized trials were lack of randomization, extremely high attrition rates, and self-reported outcomes. Risk of confounding and selection bias was common in the non-randomized studies.

Research that had less risk of bias was found to report more conservative effect sizes, indicating that some positive results in more risky research might over estimate the effectiveness of an intervention. These findings support the importance of the careful study design and clear reporting.

4.4 Discussion

Altogether, the results of this review provide the idea that AI-powered chatbots and digital applications can offer small mental health advantages to adults, especially depression and anxiety. These advantages are widely aligned with the earlier systematic reviews and meta-analyses (Vaidyam et al., 2020; Li et al., 2023), which makes AI-based interventions a potential option to support mental health in the form of a low-intensity intervention.

Nevertheless, efficacy is not the sole measure of effects in the real world. It is found that engagement is a very important moderating factor and sustained use depends on the design of the intervention, intervention structure and perceived responsiveness. Thematic synthesis

shows that completely autonomous AI systems might not be able to remain engaged over the long term without guidance or design of some sort.

The least developed area of evidence base is safety. The absence of standardized reporting adverse events and assessment of crisis management skills are some of the factors that concern the extensive implementation without more robust protection. With the increased autonomy and accessibility of AI systems, there is a need to address these safety gaps to implement AI in an ethical and responsible manner.

Clinically, AI-based mental health interventions could be better positioned as an adjunct intervention or a first-line intervention, as opposed to a substitute treatment provided by a human. Their scalability and accessibility render them appealing to prevention and early intervention, although it should be carefully coordinated with the existing care systems to handle the risk and provide proper escalation where necessary.

4.5 Future Research Implication.

The results of this review point to the relevance of future studies in the domain of mental health interventions based on AI in several directions. To begin with, there is an apparent necessity of research that creates long-term follow-up to determine the stability and sustainability of the treatment effects. Although numerous studies prove the short-term effects of the treatment on the depressive and anxiety symptoms, there is limited evidence on the maintenance of the gains in the long run. The longitudinal designs are thus necessary to find out whether AI-based interventions can lead to enduring mental health gains.

Adoption and development of standardized engagement metrics should also be a priority of future research. The existing literature is inconsistent in both the definition and measurement of engagement, and it is challenging to compare interventions. Regular reporting of compliance, intensity of use and dropout rates would make the results more interpretable and enable more rigorous synthesis of evidence.

Moreover, a safer reporting system should be mandatory and transparent, making it a fundamental part of the AI mental health research. Adverse event, symptom deterioration, crisis management consequences need systematic control and reporting to achieve ethical and responsible application of AI-based interventions. Standard safety regimes would enhance the trust in the clinical application of these technologies.

There is also a need to conduct comparative studies to compare AI-based interventions with those of therapist-led care and other available treatments. The trials would assist in determining the relative efficacy of AI-based support and guide the correct clinical placement, in particular, whether these tools are to be used as stand-alone or as supplements to conventional therapy.

Lastly, since improved generative AI systems will be more frequently deployed into mental health apps, they must be thoroughly assessed in controlled settings. A study of the advantages, constraints and dangers of these systems will become crucial towards ensuring the safe and efficient adoption of these systems. The priorities of these studies are that they will strengthen the evidence base and enable the responsible development of AI-based mental health interventions.

5. Conclusion

This systematic review included evidence that was synthesized on the effectiveness, engagement and safety of AI-based chatbots and digital tools in adult mental health. Generally, the results indicate that these interventions may yield small to moderate effect in depressive and anxiety symptoms especially in adults with mild to moderate mental health challenges. It seems that AI-based interventions seem the most effective when they are based on evidence-based

psychotherapeutic models, including cognitive behavioral therapy, and are provided in the format of structured and guided form of intervention.

Although the review indicates that the studies achieved significant results regarding user engagement, it is important to note that the studies differed significantly in how users engaged. The long-term use is the most important issue, and high dropout rates and the decrease in interaction have been noted in most interventions. Therapeutic structure, personalization, and perceived conversational quality are some of the factors that affected the engagement. These results suggest that the practical use of AI-based mental health instruments is not only limited to clinical effectiveness but also to considerate design strategies that facilitate the further participation of users.

The safety gap was a critical issue in the literature that was found. Even though the prevalence of serious adverse events was low in the studies, it can be argued that the absence of standardized safety monitoring and reports prevents concluding on risk because of inconsistency. Lack of strong testing of crisis identification, escalation policy and possible damages is even more worrying considering the autonomous character of most AI-based systems. Ethical issues pertaining to data privacy, accountability and user dependency were often considered and rarely evaluated empirically.

Combined, these sources indicate that AI-powered chatbots and digital tools can be used as useful complements to traditional mental health care, but not as substitutes to human-administered therapy. They are highly scalable and accessible and can be applied to early intervention, prevention, and supportive purposes in those facilities where professional services are scarce. Nevertheless, care should be exercised when adopting it especially in those who have severe symptoms or are at high risk.

The future research agenda should be based on high-quality randomized trials with long-term follow-up, standard measures of engagement, and compulsory reporting of safety. To make sure that AI-based interventions become an effective and ethical part of mental health care, it will be necessary to make AI system design more transparent and subjected to regulatory oversight. Through proper protection and stringent quality controls, AI-driven mental health technologies can play a valuable role in ensuring that the mental health issue is reduced in the world.

References

Andersson, G., Carlbring, P., Titov, N., & Lindefors, N. (2019). Internet-based psychological treatments: From innovation to implementation. World Psychiatry, 18(1), 20–28. https://doi.org/10.1002/wps.20610

Andersson, G., et al. (2019). Internet-based psychological treatments. World Psychiatry, 18(1), 20–28.

Andersson, G., et al. (2019). Internet-based psychological treatments: From innovation to implementation. World Psychiatry, 18(1), 20–28.

Bickmore, T. W., et al. (2018). Automated conversational agents for mental health. Journal of Medical Internet Research, 20(7), e101.

Bickmore, T. W., Trinh, H., Asadi, R., & Olafsson, S. (2018). Safety first: Conversational agents for health care. Journal of Medical Internet Research, 20(7), e101. https://doi.org/10.2196/jmir.9851

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101.

Cuijpers, P., Karyotaki, E., Reijnders, M., & Huibers, M. J. H. (2023). Who benefits from cognitive behavior therapy for depression? A meta-analytic update. World Psychiatry, 22(1), 77–89. https://doi.org/10.1002/wps.21038

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. JMIR Mental Health, 4(2), e19. https://doi.org/10.2196/mental.7785

Fitzpatrick, K. K., et al. (2017). Delivering CBT via a chatbot. JMIR Mental Health, 4(2), e19.

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

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation. JMIR mHealth and uHealth, 6(11), e12106. https://doi.org/10.2196/12106

Li, H., et al. (2023). AI conversational agents for mental health: A meta-analysis. NPJ Digital Medicine, 6, 112.

Li, H., et al. (2023). The effectiveness of artificial intelligence–based conversational agents for mental health: A meta-analysis. NPJ Digital Medicine, 6, 112.

Li, H., Zhang, R., Lee, Y. C., Kraut, R. E., & Mohr, D. C. (2023). The effectiveness of artificial intelligence–based conversational agents for mental health: Systematic review and meta-analysis. NPJ Digital Medicine, 6, 112. https://doi.org/10.1038/s41746-023-00979-5

Luxton, D. D. (2020). Ethical implications of artificial intelligence in mental health. Ethics & Behavior, 30(2), 95–109. https://doi.org/10.1080/10508422.2019.1694353

Luxton, D. D. (2020). Ethical issues in AI mental health applications. Ethics & Behavior, 30(2), 95–109.

Mohr, D. C., Lyon, A. R., Lattie, E. G., Reddy, M., & Schueller, S. M. (2021). Accelerating digital mental health research from early design and creation to successful implementation and sustainment. Journal of Medical Internet Research, 23(5), e28884. https://doi.org/10.2196/28884

Mohr, D. C., Riper, H., & Schueller, S. M. (2018). A solution-focused research approach to achieve an implementable revolution in digital mental health. JAMA Psychiatry, 75(2), 113–114. https://doi.org/10.1001/jamapsychiatry.2017.3838

Mohr, D. C., Weingardt, K. R., Reddy, M., & Schueller, S. M. (2017). Three problems with current digital mental health research… and three things we can do about them. Psychiatric Services, 68(5), 427–429. https://doi.org/10.1176/appi.ps.201600541

Page, M. J., et al. (2021). The PRISMA 2020 statement. BMJ, 372, n71.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Sterne, J. A. C., et al. (2016). ROBINS-I. BMJ, 355, i4919.

Sterne, J. A. C., et al. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. BMJ, 355, i4919.

Sterne, J. A. C., et al. (2019). RoB 2. BMJ, 366, l4898.

Sterne, J. A. C., et al. (2019). RoB 2: A revised tool for assessing risk of bias in randomized trials. BMJ, 366, l4898.

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., … Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. BMJ, 355, i4919. https://doi.org/10.1136/bmj.i4919

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomized trials. BMJ, 366, l4898. https://doi.org/10.1136/bmj.l4898

Torous, J., et al. (2020). Digital mental health engagement challenges. NPJ Digital Medicine, 3, 110.

Torous, J., Myrick, K. J., Rauseo-Ricupero, N., & Firth, J. (2020). Digital mental health and COVID-19: Using technology today to accelerate the curve on access and quality tomorrow. JMIR Mental Health, 7(3), e18848. https://doi.org/10.2196/18848

Torous, J., Nicholas, J., Larsen, M. E., Firth, J., & Christensen, H. (2018). Clinical review of user engagement with mental health smartphone apps. Journal of Medical Internet Research, 20(7), e214. https://doi.org/10.2196/jmir.8852

Vaidyam, A. N., et al. (2020). Conversational agents in mental health. Journal of Medical Internet Research, 22(8), e16868.

Vaidyam, A. N., et al. (2020). Conversational agents in mental health. Journal of Medical Internet Research, 22(8), e16868.

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2020). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. Journal of Medical Internet Research, 22(8), e16868. https://doi.org/10.2196/16868

World Health Organization. (2022). World mental health report: Transforming mental health for all. WHO. https://www.who.int/publications/i/item/9789240049338