Dean&Francis

ISSN 2959-6122

AL004642

The effectiveness of chatbots and its difference with usual treatment to reduce the symptoms of depression and anxiety: meta-analysis review

Ge Feng*

Guangzhou Zhixin High School, 510199, China

**Corresponding author email: anthonyfeng99@foxmail.com*

Abstract:

The review aims to test the effectiveness of chatbots in treating depression and anxiety through cognitive behavioral therapy within a short duration of treatment. This review also tries to test whether chatbot therapies are similar to usual treatments. The study pooled data from 12 studies. In the random effects model, at the significance level of α = 0.05, chatbot therapies significantly reduce the symptoms of depression and anxiety with a small to moderate effect size. For anxiety groups, the range of summarized standardized mean difference is from -0.39 to -0.1. For depression groups, the interval of a summarized standardized mean difference ranges from -0.36 to -0.08. The difference in effectiveness between chatbot therapies and traditional therapies is minimal. Therefore, chatbot-provided cognitive behavioral therapy shows promise for the future. Three studies with significant heterogeneity were removed. They differ in terms of outdated chatbots, significantly varying durations of treatments, and the special period during which the study was conducted. These implications may suggest that chatbots are undergoing rapid upgrades, so future researchers should keep up with the changes in chatbot technology. The number of related researches is limited, and moral concern should be considered in future researches.

Keywords: chatbots, depression, anxiety, CBT, meta-analysis

1. Introduction

Conversational artificial intelligence (AI-based chatbot) is prevalent. Patel et al. (2019) tried to establish several AI networks to understand human emotions and identify them as labels including but not limited to joy, pain, and stress. The result shows that AI can comprehend emotions in its own way, so it is possible to let chatbots take part in tasks requiring an understanding of people's emotions. Vaidyam et al. (2019) show that preliminary evidence has shown that chatbots are favorable for psychiatric use despite the lack of further research. In the study, one noticeable point is that patients are inclined to positively rate treatments provided by chatbots, suggesting that chatbots have the potential to offer enjoyable medical experiences for patients. Olawade et al. (2024) suggest that chatbots can significantly change the way of mental healthcare, but privacy and bias are noticeable and need to be considered in the future.

Two of the most significant mental disorders are depression and anxiety. Major depressive disorder can lead to severe situations. Patients diagnosed with depression disorder may have low mood, hopelessness, and insomnia. Family or friend relationships may be affected. Generalized anxiety disorder, on the other hand, makes the diagnosed patients respond in overreaction. The patients are excessively worried, significantly disturbing their lives.

In 2023, WHO (2023a) (World Health Organization) surveys that anxiety is affecting 302 million people. Another article from WHO (2023b) points out that 5% of adults are suffering from depression. These results suggest that depression and anxiety are highly prevalent.

Symptoms of depression and anxiety can lead to suicide attempts, so the prevalence and severity show that depression and anxiety need to be considered. Young et al. (2008) suggests nearly 30% people diagnosed with depression or anxiety can receive treatments and cures. Therefore, it is necessary to find an approach to make respective cures more accessible.

Therefore, we will focus on the topic of generalized anxiety disorders and major depressive disorders. They are highly influential and relatively prevalent. One of the reasons for this may be the shortage of therapists. Boucher et al. (2021) suggest that chatbots may cover the shortage of practitioners in the psychological field despite the lack of further research on distinguishing the effectiveness between chatbots and other digital interventions. Also, Zhong et al. (2024) show that chatbots are cost-effective and highly accessible. Moreover, chatbots can provide cognitive behavioral therapy via communicating with patients. Cognitive behavioral therapies have been proven to have high effectiveness in treating mental disorders including depression and anxiety.

Although Abd-Alrazaq et al. (2020) point out that the evidence proving that chatbots are capable of curing mental disorders is insufficient, recently, Zhong et al. (2024) conduct a

meta-analysis to conclude that chatbots can relieve the symptoms of depression and anxiety via cognitive behavioral therapy. Granted, there are limitations including a lack of variety in chatbots taken into account. Also, they point out that the number of existing studies is still limited. Despite the limitation of studies and data sources, we try to add the diversity of chatbots in one

meta-analysis, figuring out if conversational artificial intelligence is truly effective in alleviating the symptoms of depression and anxiety, instead of a bias produced by specific chatbots. This is important to test because we need to clarify if the reduction of severity is caused by specific types of chatbots with certain confounding variables.

If they are, is the effectiveness of the therapies provided by chatbots different from traditional therapies? We conduct t-tests of equivalence to find the implication.

To validate our hypothesis, we conduct meta-analysis using statistical tools. Despite the heterogeneity can affect the analysis, Higgins and Thompson (2002) suggest that heterogeneity can be well measured by several indicators including I^2. In order to validate our meta-analysis, we use I^2 as well as χ^2 to detect the heterogeneity and remove studies that significantly affect these indicators.

2. Method

We hypothesize that chatbot therapies are effective and similar to artificial treatment. To validate the hypotheses, we conduct testing toward these two hypotheses separately.

We use the method in Kumar et al. (2022). To test whether or not the chatbots therapies are effective, during this procedure, the calculation will focus on PHQ-9 and GAD-7 scores of chatbot groups (patients receiving CBT from chatbots) and non-intervention groups (patients receiving no treatment or a psychological book. We calculate the standardized mean difference of PHQ-9 and GAD-7 between chatbot groups and non-intervention groups. The estimation of the standardized mean difference is Cohen's d. In order to test the null hypothesis, we build the confidence interval of the effect size Cohen's d, which indicates how significant the difference is.

Based on Lin and Chu (2018), in order to test if there is publication bias, which can highly affect the conclusion of meta-analysis, we use Egger's test and funnel plot to detect potential publication bias, increasing the reliability of the result.

We verify if the chatbot therapies are significantly different from traditional or artificial therapies. To this end, we utilize t-tests of equivalence test for two independent samples. The groups being tested used the same measurement of severity of depression and anxiety symptoms, which are PHQ-9 and GAD-7. To ensure the consistency of this meta-analysis, heterogeneity is tested in every process by I^2 and χ^2. During the process of data extraction, significant outliers are removed.

2.1 Participants

We collect 36 studies from 11 papers with a total number of 2173 subjects within control and experimental groups. The following categories of data are not tractable in every paper collected. Therefore, the estimated distributions of demography are approximate. The mean of the age of subjects is 30.950(7.135). 75.9% of subjects are women. 24.0% of subjects are males. Nonbinary sexuality is too small in the collection to account. 42.9% of subjects have received more than 12th grade.

Chatbot groups versus non-treatment groups the null hypothesis is that PHQ-9 and GAD-7 scores of chatbot groups are equal with the groups receiving no treatment. To test the hypothesis, this paper calculates the PHQ-9 and GAD-7 scores separately.

2.2 Materials

Data in the following studies is collected. We vary the chatbots in this review. The studies signed with star are removed because they significantly affect heterogeneity. In Liu et al. (2022)'s study, the treatment conducts for 16 weeks, differing from other studies' duration significantly. Fitzpatrick et al. (2017)'s study conducted at 2017, so the implication may tell us that the heterogeneity is produced by the outdated chatbot, and chatbots are rapidly progressive. He et al. (2022)'s study is conducted during the Covid-19 pandemic. This may imply that the environment will influence mental health, affecting the effectiveness of the therapies. (Table 1).

Table 1

Study	AI	Duration(weeks)
Sadeh-Sharvit et al., 2023	The Eleos Health Platform	4,8
Danieli et al., 2022	TEO	4,8
Klos et al., 2021	Tess	8
Fitzpatrick et al., 2017*	Woebot	2
Liu et al., 2022*	XiaoNan	16
Nicol et al., 2022	Mobile Health	4
Karkosz et al., 2024	Fido	2
Suharwardy et al., 2023	Woebot	6
He et al., 2022*	XiaoE	4
Ulrich et al., 2024	MISHA	4,7(mixed)
Greer et al., 2019	Vivibot	4
Anmella et al., 2023	Vickybot	2,4
Suharwardy et al., 2023	Woebot	6
Guțu et al., 2021	Woebot	2
MacNeill et al., 2024	Wysa	2,4

2.3 Measures

To quantitatively measure the symptoms of depression and anxiety, we mainly look for trials using the measurement of GAD-7 and PHQ-9 scores.

2.3.1 GAD-7 Generalized anxiety disorder-7

According to Löwe et al. (2008), GAD-7 is a reliable questionnaire that can measure anxiety in the general population. It has been used in a wide range of clinical usage. Commonly, a score between 10 and 14 would be considered moderate anxiety, and a score between 15 and 19 is severe anxiety.

2.3.2 PHQ-9 Patient health questionnaire-9

According to Sun et al. (2020), similarly, PHQ-9 is a specific questionnaire that quantitatively measures the severity of major depressive disorders (clinical depression). It has also been used to apply to the giant range. A score between 10 and 14 would be considered moderate anxiety, and a score between 15 and 19 is moderately severe. A score above 19 is severe depression.

2.4 Procedure

According to the figure 1, based on databases (google scholar), we search for studies conducting randomized controlled trials. Typically, studies should measure the severity in both controlled groups and groups who receive CBT provided by chatbots. In order to compare the effectiveness between chatbots and non-treatment groups, the controlled groups will be patients receiving non-treatment or just a psychological handbook. On the other hand, to test if chatbot treatments are similar to usual treatments, the controlled groups will receive treatments as usual. To quantitatively measure the severity of symptoms and maintain consistency, we try to prioritize studies using the measurements of PHQ-9 and GAD-7. Significant outliers, which make the heterogeneity extremely high, are removed.

3. Results Chatbot groups versus non-treatment

3.1 groups PHQ-9

According to figure 2, all together 13 studies were analyzed with a total of 362 subjects in the experimental cohort and 472 subjects in the control cohort. There are 8 different kinds of chatbots within 13 studies, ensuring that the conclusion is consistent in various chatbots.

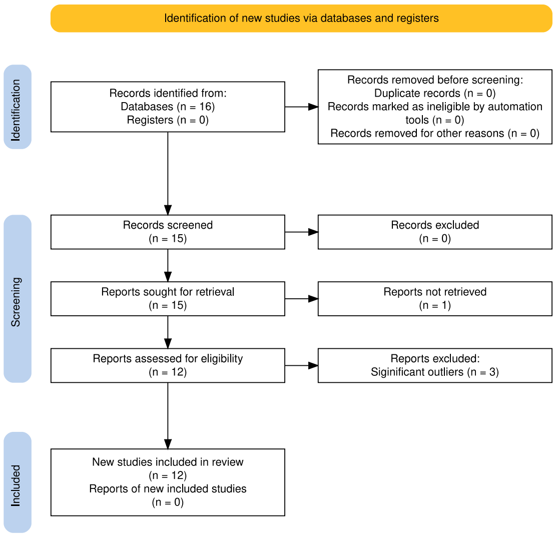
Identification of new studies via databases and registers
Identification
Records identified from:
Databases (n = 16)
Registers (n = 0)
Records removed before screening:
Duplicate records (n = 0)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 0)
Screening
Records screened
(n = 15)
Records excluded
(n = 0)
Reports sought for retrieval
(n = 15)
Reports not retrieved
(n = 1)
Reports assessed for eligibility
(n = 12)
Reports excluded:
Significant outliers (n = 3)
Included
New studies included in review
(n = 12)
Reports of new included studies
(n = 0)

Figure 1

The PHQ-9 scores of chatbot' therapies are significantly less than the scores of patients receiving no treatment. Based on the analysis performed using the random effects model with the Inverse variance method to compare the standardized mean difference (SMD), there is a statistical difference between the two cohorts, the summarized standardized mean difference (SMD) is -0.22 with a 95% confidence interval of $(-0.36, -0.08)$. According to Kumar et al., 2022, the difference is significant and has perhaps a small and moderate effect. The test for overall effect shows a significance at $p < 0.05$.

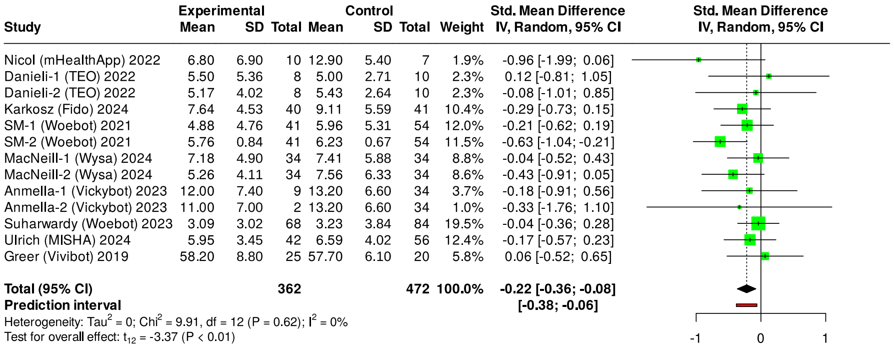

Figure 2

Significant heterogeneity was not observed, indicating that effect sizes across studies are consistent in both magnitude and direction.

3.2 GAD-7

According to figure 3, all together 13 studies were analyzed with a total of 333 subjects in the Experimental cohort and 422 subjects in the Control cohort. There are 9 different kinds of chatbots within 13 studies, ensuring that the conclusion is consistent in various chatbots.

The GAD-7 scores of chatbot therapies are significantly less than the scores of patients receiving no treatment. Based on the analysis performed using the random effects model with the Inverse variance method to compare the standardized mean difference (SMD), there is a statistical difference between the two cohorts, the summarized standardized mean difference (SMD) is -0.24 with a 95% confidence interval of $(-0.39, -0.1)$. According to Kumar et al., 2022, the difference is significant and has perhaps a small and moderate effect. The test for overall effect shows a significance at $p < 0.05$.

Significant heterogeneity was not observed, indicating that effect sizes across studies are consistent in both magnitude and direction.

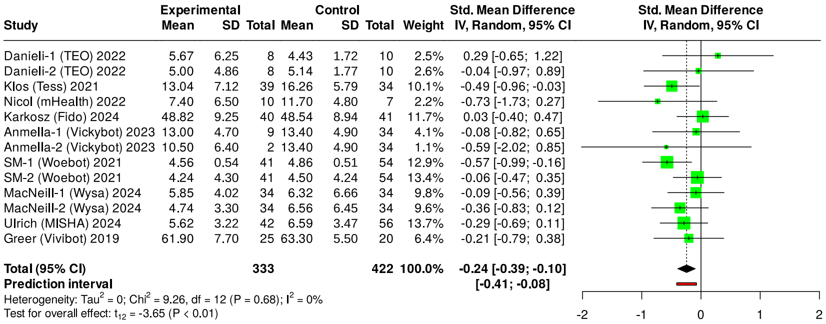

Figure 3 GAD-7

3.3 Chatbot groups versus treatment as usual

The null hypothesis is that the difference between the scores of chatbot groups and the groups receiving usual treatment is less than 2 and more than -2.

To test these hypotheses, this paper calculates the PHQ-9 and GAD-7 scores separately.

3.3.1 PHQ-9

All together 5 studies were analyzed with a total of 130 subjects in the Experimental

cohort and 162 subjects in the Control cohort. There are 3 different types of chatbots within 5 studies. The null hypothesis is equivalence to:

$H_0 = |\mu AI - \mu TAU| > 2$ (1)

We need to test two null hypotheses.

$H_0 = \mathrm{\mu} AI - \mu TAU > 2$ (2)

$t = -2.422$, $d\,f = 206.732$, $p = .008$. The one-tail null hypothesis is rejected at the significant level α = 0.05.

$H_0 = \mathrm{\mu} AI - \mu TAU > -2$ (3)$t = 1.939$, $d\,f = 206.732$, $p = .028$. The one-tail null hypothesis is rejected at the significant level

α = 0.05.

There is convincing statistical evidence that the absolute difference in the PHQ-9 scores between the groups of chatbots and those receiving treatment as usual is less than or equal to 2.

By testing, ($\chi^2 = 4.26$, $d\,f = 4$, $p = 0.37$; $I^2 = 6\%$), significant heterogeneity was not observed,

indicating that effect sizes across studies are consistent in both magnitude and direction.

3.3.2 GAD-7

All together 5 studies were analyzed with a total of 130 subjects in the Experimental cohort and 162 subjects in the Control cohort. There are 3 different types of chatbots within 5 studies. The null hypothesis is:

We need to test two null hypotheses.

$H_0 = |\mu AI - \mu TAU| > 2$ (4)$H_0 = \mathrm{\mu} AI - \mu TAU > 2$ (5)

$t = -2.227$, $d\,f = 215.43$, $p = 0.014$. The one-tail null hypothesis is rejected at the significant level α = 0.05.

$H_0 = \mathrm{\mu} AI - \mu TAU > -2$ (6)

$t = 1.707$, $d\,f = 215.43$, $p = 0.045$. The one-tail null hypothesis is rejected at the significant level

α = 0.05.

There is convincing statistical evidence that the absolute difference in the GAD-7 scores between the groups of chatbots and those receiving treatment as usual is less than or equal to 2. By testing, $\chi 2 = 2.8$, $d\,f = 4$, $p = 0.59$; $I^2 = 0\%$, significant heterogeneity was not observed, indicating that effect sizes across studies are consistent in both magnitude and direction.

4. Discussion

We can make an inference from the statistics above. Because the entire range of Cohen's d is negative, meaning that the PHQ-9 or GAD-7 scores of chatbot groups are less than controlled groups, Chatbots can cure patients diagnosed with major depressive disorder or generalized anxiety disorders, and there is statistically convincing evidence showing that the treatments can be provided by well-developed conversational artificial intelligence. Therefore, chatbots are showing a promising future in the clinical field. By varying the types of chatbots, we can ensure that patients whose severity of mental disorders gets reduced not because specific types of chatbots have confounding variables. From the statistical analysis, we have taken various types of chatbots into account, so the conclusion can probably be generalized to a wider range of chatbots instead of specific types.

According to the t-tests, we are confident to confirm that the difference between chatbot treatments and usual treatments is less than 2 scores in both PHQ-9 and GAD-7. Considering each level of severity in these two questionnaires is 4, we consider that 2 scores difference may not be huge enough to affect the symptoms in real situations. Also, chatbots have multiple advantages including their cost-effectiveness and high accessibility.

Consequently, we think that chatbots can help cure patients with depression or anxiety disorders. Future studies can try to enhance the ability to deal with severe mental disorders and find an approach to widen the usage of chatbots through applications, increasing its accessibility to more patients who cannot receive adequate treatments yet. However, current studies are still limited. We need more experiments and data to confirm the hypothesis, enhancing the reliability of the conclusions. Also, the duration of treatments should be prolonged. Studies excluded, which are identified to significantly affect heterogeneity, set a huge difference in duration of treatments compared with other studies included. Liu et al. (2022) 's study, which causes giant heterogeneity in these studies, set a treatment for 16 weeks, significantly differing from other groups included, which average set 4 to 6 weeks. Therefore, it is possible that the heterogeneity is produced by the significant difference in duration. However, existing studies are rare to set long enough experiments, so this meta-analysis fails to consider long-term therapies.

Chatbots are experiencing rapid upgrading. From previous statistical tests, chatbots between even several years can produce significant heterogeneity, so it is important to collectively conduct a huge experiment containing giant data to fully validate the hypothesis.

There are concerns on chatbots. Although chatbots probably are effective in the macro, misconducting is dangerous to patients, and it is uncertain who can be held liable for a miscarriage of justice or mistreatment. Therefore, the wide range of usage of chatbots is difficult to come true without eliminating moral concerns.

References

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.

Anmella, G., Sanabra, M., Primé-Tous, M., Segú, X., Cavero, M., Morilla, I., Grande, I., Ruiz, V., Mas, A., Martín-Villalba, I., et al. (2023). Vickybot, a chatbot for anxiety-depressive symptoms and work-related burnout in primary care and health care professionals: Development, feasibility, and potential effectiveness studies. Journal of medical Internet research, 25, e43293.

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

Danieli, M., Ciulli, T., Mousavi, S. M., Silvestri, G., Barbato, S., Di Natale, L., & Riccardi, G. (2022). Assessing the impact of conversational artificial intelligence in the treatment of stress and anxiety in aging adults: Randomized controlled trial. JMIR mental health, 9(9), e38067.

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), e7785.

Greer, S., Ramo, D., Chang, Y.-J., Fu, M., Moskowitz, J., Haritatos, J., et al. (2019). Use of the chatbot "vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: Randomized controlled feasibility trial. JMIR mHealth and uHealth, 7(10), e15018.

Gut,u S., Cosmoiu, A., Cojocaru, D., Turturescu, T., Popoviciu, C., & Giosan, C. (2021). Bot to the rescue? effects of a fully automated conversational agent on anxiety and depression: A randomized controlled trial. Ann Depress Anxiety, 8(1), 1107.

He, Y., Yang, L., Zhu, X., Wu, B., Zhang, S., Qian, C., & Tian, T. (2022). Mental health chatbot for young adults with depressive symptoms during the covid-19 pandemic: Single-blind, three-arm randomized controlled trial. Journal of medical Internet research, 24(11), e40719.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in medicine, 21(11), 1539–1558.

Karkosz, S., Szyman´ski, R., Sanna, K., Michałowski, J., et al. (2024). Effectiveness of a

web-based and mobile therapy chatbot on anxiety and depressive symptoms in subclinical young adults: Randomized controlled trial. JMIR formative research, 8(1), e47960.

Klos, M. C., Escoredo, M., Joerin, A., Lemos, V. N., Rauws, M., & Bunge, E. L. (2021). Artificial intelligence–based chatbot for anxiety and depression in university students: Pilot randomized controlled trial. JMIR formative research, 5(8), e20678.

Kumar, L. M., Stephen, J., George, R., Harikrishna, G., & Anisha, P. (2022). Use of effect size in medical research: A brief primer on its why and how. Kerala Journal of Psychiatry, 35(1), 78–82.

Lin, L., & Chu, H. (2018). Quantifying publication bias in meta-analysis. Biometrics, 74(3), 785–794.

Liu, H., Peng, H., Song, X., Xu, C., & Zhang, M. (2022). Using ai chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness.

Internet Interventions, 27, 100495.

Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., & Herzberg, P. Y. (2008). Validation and standardization of the generalized anxiety disorder screener (gad-7) in the general population. Medical care, 46(3), 266–274.

MacNeill, A. L., Doucet, S., & Luke, A. (2024). Effectiveness of a mental health chatbot for people with chronic diseases:

Randomized controlled trial. JMIR Formative Research, 8, e50025.

Nicol, G., Wang, R., Graham, S., Dodd, S., Garbutt, J., et al. (2022). Chatbot-delivered cognitive behavioral therapy in adolescents with depression and anxiety during the covid-19 pandemic: Feasibility and acceptability study. JMIR Formative Research, 6(11), e40242.

Olawade, D. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024). Enhancing mental health with artificial intelligence: Current trends and future prospects. Journal of Medicine, Surgery, and Public Health, 3, 100099. https://doi.org/https://doi.org/10.1016/j.glmedi.2024.100099

Patel, F., Thakore, R., Nandwani, I., & Bharti, S. K. (2019). Combating depression in students using an intelligent chatbot: A cognitive behavioral therapy. 2019 IEEE 16th India Council International Conference (INDICON), 1–4.

Sadeh-Sharvit, S., Camp, T. D., Horton, S. E., Hefner, J. D., Berry, J. M., Grossman, E., & Hollon, S. D. (2023). Effects of an artificial intelligence platform for behavioral interventions on depression and anxiety symptoms: Randomized clinical trial. Journal of Medical Internet Research, 25, e46781.

Suharwardy, S., Ramachandran, M., Leonard, S. A., Gunaseelan, A., Lyell, D. J., Darcy, A., Robinson, A., & Judy, A. (2023). Feasibility and impact of a mental health chatbot on postpartum mental health: A randomized controlled trial. AJOG global reports, 3(3), 100165.

Sun, Y., Fu, Z., Bo, Q., Mao, Z., Ma, X., & Wang, C. (2020). The reliability and validity of phq-9 in patients with major depressive disorder in psychiatric hospital. BMC psychiatry, 20, 1–7.

Ulrich, S., Lienhard, N., Künzli, H., & Kowatsch, T. (2024). A chatbot-delivered stress management coaching for students (misha app): Pilot randomized controlled trial. JMIR mHealth and uHealth, 12, e54945.

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(7), 456–464.

WHO. (2023a). Anxiety disorders.

https://www.who.int/news-room/fact-sheets/detail/anxiety-disorders WHO. (2023b). Depressive disorder.

https://www.who.int/news-room/fact-sheets/detail/depression

Young, A. S., Klap, R., Shoai, R., & Wells, K. B. (2008). Persistent depression and anxiety in the united states: Prevalence and quality of care. Psychiatric Services, 59(12), 1391–1398.

Zhong, W., Luo, J., & Zhang, H. (2024). The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: A systematic review and meta-analysis. Journal of Affective Disorders.