Journal Pre-proof

Feasibility and impact of a mental health chatbot on postpartum mental health: A randomized controlled trial

Sanaa Suharwardy , Maya Ramachandran , Stephanie A. Leonard , Anita Gunaseelan , Deirdre J. Lyell , Alison Darcy , Athena Robinson , Amy Judy

PII: S2666-5778(23)00006-0
DOI: https://doi.org/10.1016/j.xagr.2023.100165
Reference: XAGR 100165

To appear in: *AJOG Global Reports*

Received date: 19 July 2022
Revised date: 19 December 2022
Accepted date: 22 January 2023

Please cite this article as: Sanaa Suharwardy , Maya Ramachandran , Stephanie A. Leonard , Anita Gunaseelan , Deirdre J. Lyell , Alison Darcy , Athena Robinson , Amy Judy , Feasibility and impact of a mental health chatbot on postpartum mental health: A randomized controlled trial, *AJOG Global Reports* (2023), doi: https://doi.org/10.1016/j.xagr.2023.100165

Title: Feasibility and impact of a mental health chatbot on postpartum mental health: A randomized controlled trial

Authors: Sanaa Suharwardy[1], Maya Ramachandran[1], Stephanie A. Leonard[1], Anita Gunaseelan[1], Deirdre J. Lyell[1], Alison Darcy[2], Athena Robinson[2], Amy Judy[1]

[1]Stanford University, Stanford, CA

[2]Woebot Health, San Francisco, CA

Patient consent is not required because no personal information or details are included.

Corresponding Author: Sanaa Suharwardy, MD, Center for Academic Medicine, Obstetrics and Gynecology MC 5317, Stanford University School of Medicine, 453 Quarry Road, Palo Alto, CA, USA 94304-5317. Phone: (650) 736-1191. Fax: (650) 723-7737. Email: sanaas@stanford.edu

Disclosures: Alison Darcy, PhD and Athena Robinson, PhD are Woebot Health employees. The remaining authors report no conflict of interest.

Trial Registration: ClinicalTrials.Gov NCT03646539

https://clinicaltrials.gov/ct2/show/NCT03646539

Condensation: Use of a smartphone application-based mental health chatbot for mood management demonstrated feasibility and acceptability among a general postpartum population.

Short Title: Randomized controlled trial of a postpartum mental health chatbot.

AJOG at a Glance:

A. Why was this study conducted? To pilot the acceptability and preliminary efficacy of a smartphone application-based mental health chatbot for postpartum mental health.

B. What are the key findings? Participants randomized to use the chatbot reported high satisfaction with the chatbot and found it highly acceptable as a mental health management tool. Among a general postpartum and non-depressed population, the use of a smartphone application-based mental health chatbot for the first 6 weeks postpartum reduced depression scores from baseline on one assessment tool that is commonly used in clinical practice.

C. What does this study add to what is already known? This randomized controlled trial demonstrates that a digital mental health resource, such as a chatbot, may contribute to mood

management and improvement in the postpartum period, and that such approaches are of interest and acceptable to patients.

Keywords: postpartum depression, perinatal mood, telehealth, digital therapeutics, artificial intelligence, chatbot, smartphone app, mental health

Acknowledgements: The authors would like to thank Dr. Anna Girsen and Angela Clear for their support and guidance. This work was funded by grants from the Stanford Medical Scholars, Stanford Society of Physician Scholars, and Stanford Maternal Child Health Research Institute.

ABSTRACT

Background: Perinatal mood disorders are common yet underdiagnosed and un- or undertreated. Barriers exist to accessing perinatal mental health services, including limited availability, time, and cost. Automated conversational agents (chatbots) can deliver evidence-based cognitive behavioral therapy content through text message-based conversations and reduce depression and anxiety symptoms in select populations. Such digital mental health technologies are poised to overcome barriers to mental health care access but need to be evaluated for efficacy, as well as for preliminary feasibility and acceptability among perinatal populations.

Objective: To evaluate the acceptability and preliminary efficacy of a mental health chatbot for mood management in a general postpartum population.

Study Design: An unblinded randomized controlled trial was conducted at a tertiary academic center. English-speaking postpartum women aged 18 years or above with a live birth and access to a smartphone were eligible for enrollment prior to discharge from delivery hospitalization. Baseline surveys were administered to all participants prior to randomization to a mental health chatbot intervention or to usual care only. The intervention group downloaded the mental health chatbot smartphone application with perinatal-specific content, in addition to continuing usual care. Usual care consisted of routine postpartum follow up and mental health care as dictated by the patient's obstetric provider. Surveys were administered during delivery hospitalization (baseline) and at 2-, 4-, and 6-weeks postpartum to assess depression and anxiety symptoms. The primary outcome was a change in depression symptoms at 6-weeks as measured using two depression screening tools: PHQ (Patient Health Questionnaire-9) and EPDS (Edinburgh Postnatal Depression Scale). Secondary outcomes included anxiety symptoms measured using GAD (Generalized Anxiety Disorder-7), and satisfaction and acceptability using validated scales. Based on a prior study, we estimated a sample size of 130 would have sufficient (80%) power to detect a moderate effect size (d=.4) in between group difference on the PHQ-9.

Results: 192 women were randomized equally 1:1 to the chatbot or usual care; of these, 152 women completed the 6-week survey (n=68 chatbot, n=84 usual care) and were included in the

final analysis. Mean baseline mental health assessment scores were below positive screening thresholds. At 6-weeks, there was a greater decrease in PHQ-9 scores among the chatbot group compared to the usual care group (mean decrease=1.32, SD=3.4 vs mean decrease=0.13, SD=3.01, respectively). 6-week mean EPDS and GAD-7 scores did not differ between groups and were similar to baseline. 91% (n=62) of the chatbot users were satisfied or highly satisfied with the chatbot, and 74% (n=50) of the intervention group reported use of the chatbot at least once in 2 weeks prior to the 6-week survey. 80% of study participants reported being comfortable with the use of a mobile smartphone application for mood management.

Conclusion: Use of a chatbot was acceptable to women in the early postpartum period. The sample did not screen positive for depression at baseline and thus the potential of the chatbot to reduce depressive symptoms in this population was limited. This study was conducted in a general obstetric population. Future studies of longer duration in high-risk postpartum populations who screen positive for depression are needed to further understand the utility and efficacy of such digital therapeutics for that population.

INTRODUCTION

Childbirth is a significant event, accompanied by definitive life changes, including physical, behavioral, psychological, and emotional adjustments. The American College of Obstetricians and Gynecologists (ACOG) recommends screening of all women for depression and anxiety symptoms at the comprehensive postpartum visit. Worldwide, about 10-20% of women develop

the full clinical syndrome of peripartum depression,[1] defined as minor or major depressive episodes with onset during pregnancy or within the first 12 months following childbirth.[2] The American Psychiatric Association's (APA) guidelines for the treatment of women with major depression who are pregnant or breastfeeding indicate psychotherapy without medication as a first line treatment.[3] Cognitive behavioral therapy (CBT) and interpersonal psychotherapy (IPT) are evidence-supported and recommended psychotherapies for PPD.[4-7]

Research on cognitive behavioral therapy translated into digital forms has demonstrated efficacy in reducing symptoms of depression with enormous potential to scale-up access.[4-6] Digital adaptations continue to evolve, and newer applications have focused on conversational agents, software programs that use artificial intelligence to simulate a conversation with a user through written text or voice.[9] Smartphone ownership is at an all-time high in the United States, with 96% of 18–29-year-olds and 95% of 30–49-year-olds owning a smartphone. The rates were similar across ethnicities, which may help to reduce health care disparities in access.[10] One currently available smartphone application offers an instantaneously available text-based conversational agent, or "chatbot," which 'checks in' with users. Using conversational tones, the chatbot encourages mood tracking and delivers perinatal-specific psychoeducation as well as tailored empathy, behavioral pattern insight, and cognitive behavioral therapeutic elements.

In a previous randomized controlled trial in adolescents, the application yielded high engagement and significant reductions in depressive symptoms compared to a control group

among young adults11 and adolescents [18] as well as substance use reductions among adults. [17] Our objective was to assess the acceptability and preliminary efficacy of a mental health chatbot on mood management in a general postpartum population.

MATERIALS AND METHODS

This was a single center, nonblinded randomized controlled trial conducted at an academic institution comparing mental health assessment scores between intervention and control groups. The Stanford Research Compliance Office approved this study. The study was registered at clinicaltrials.gov (NCI: NCT03646539) and the CONSORT guidelines were followed.[12]

Participants

Women were recruited within 72 hours of their delivery, between January 2019 and May 2019 at the Lucile Packard Children's Hospital at Stanford. Inclusion criteria were English-speaking, being 18 years or older, and access to a smartphone. Exclusion criteria were a fetal or neonatal demise (see Figure 1, enrollment diagram). After giving informed consent, participants completed a baseline survey. Randomization was performed via block randomization with a 1 to 1 allocation to usual care or use of the chatbot plus usual care. All participants received a $25 gift card after completion of the baseline survey and after completion of the 6-week survey.

Control: Usual Care

Usual care consisted of usual postpartum management at the discretion of the patient's obstetric provider. At the study site, Lucile Packard Children's Hospital, patients are generally seen for an in-person obstetric visit at 6-weeks postpartum wherein an Edinburgh Postnatal Depression Scale (EPDS) is self-administered. Some patients may also have an earlier in-person or telehealth postpartum phone visit depending on their provider and their specific comorbidities.

Intervention: Chatbot and Usual Care

The chatbot, called Woebot, was developed by Woebot Health and has demonstrated reductions in symptoms of depression among a young adult population[11] and adolescents[18] as well as substance use reductions among adults [17] For this study, peripartum-specific content was adapted to support postpartum mothers managing mood and anxiety and added to Woebot. The program delivers psychotherapeutic techniques derived from CBT and IPT for postpartum mood and invites users to track and discover patterns in their mood through text-based conversations in the smartphone application (Figure 2, Woebot picture). The intervention group was given instructions on how to download the Woebot smartphone application and access the perinatal content. Study participants were encouraged to use the program daily with the opportunity to opt-out of daily check-ins, in addition to usual care by their obstetric care team. A detailed description of a Woebot program for diagnosed postpartum depression, an adjunct to clinical supervised outpatient care management, is available via a published review

by Darcy et al. [21] Hospital course, management, and discharge timing were not affected by study participation.

Study participants were informed, through both app-onboarding procedures as well as through study informed consent, that Woebot was not designed to thoroughly assess nor intervene in suicidal ideation or other crises. Woebot is not a crisis/emergency service; this is communicated to study participants at informed consent and during the application onboarding process, at the initiation screen during application use, as well as any time the "Language Detection" Protocol (LDP) is used. [21] Woebot's LDP uses natural language processing algorithms to detect and flag several hundred potential self-harm phrases (including some misspellings and slang phrases) with 98% accuracy (sensitivity=97% and specificity=99%; Woebot Health unpublished data, September 2020). The purpose of the LDP is to detect concerning topics within patient-input free-text. Upon detection and confirmation of any concerning topics, LDP initiates a conversation between the chatbot Woebot and the participant to remind them of the application's limitations of services and offer a resource list which includes readily accessible support channels, (e.g., 9-1-1 for Emergency Services, suicide crisis hotlines), curated from consultation with experts in mental health.

Surveys

Study assessments were self-reported online via Qualtrics surveys at baseline and 2-, 4-, and 6-weeks. Baseline surveys asked about i. demographic characteristics, ii. psychiatric history, iii.

opinions regarding psychiatric care, and iv. depression and anxiety. Sample demographic characteristics were age, race/ethnicity, and marital and employment status. Psychiatric history questions asked about pre-existing mental health conditions, previous and current use of psychotherapy and psychotropic medication and had binary (yes or no) answer choices. Opinions regarding psychiatric care during the postpartum period (see Table 1) asked about current use of smartphone applications to support mental health, comfort with smartphone applications for mood management, satisfaction with usual care for depression and anxiety during the postpartum period, the importance of monitoring depression and anxiety throughout one's life and during the postpartum period, and stigma faced by women who seek mental health services postpartum. These questions about opinions regarding psychiatric care were asked using a 5-point Likert scale.

Depression and anxiety were assessed using the Edinburgh Postnatal Depression Scale (EPDS), a validated screening for depressive symptomatology among postpartum women; the 9-item Patient Health Questionnaire (PHQ-9), a widely used self-report measure of depression symptoms with demonstrated reliability and sensitivity to clinical change; and the 7-item Generalized Anxiety Disorder scale (GAD-7), a self-report tool to assess the frequency and severity of anxious thoughts and behaviors over the past 2 weeks. The 6-week survey asked about depression and anxiety using these same measures. The PHQ-9 and the EPDS were selected for use in the current study because they both are i. supported by meta-analytic data as gold-standard and psychometrically reliable and valid screening measures for depression;[2, 22]

ii. recommended by ACOG as screening instruments[2]; iii. easily administered and brief; and iv. Commonly utilized, in real world clinical settings. [23]

In addition, participants in the chatbot group were asked at 6 weeks postpartum about their i. satisfaction with the chatbot using the Client Satisfaction Questionnaire (CSQ-8), an 8 item measure of client satisfaction with mental health services.; ii. therapeutic relationship with chatbot using the Working Alliance Inventory Short -Revised version (WAI-SR); and iii. Frequency of chatbot use over the previous 2-weeks. Participants in the chatbot group were also asked for comments about their experience.

The Working Alliance Inventory-Short Revised (WAI-SR) is a measure of the therapeutic alliance, the therapeutic relationship between provider and patient that is comprised of three key aspects which are important to the practice of therapy and measured in the WAI-SR. These three aspects are: (a) agreement on the tasks of therapy, (b) agreement on the goals of therapy, and (c) development of an affective bond between provider and patient. Therapeutic alliance between Woebot and program users was previous demonstrated among a general population of over 30,000 adults.[20] Therapeutic alliance is associated with the efficacy of psychotherapy.[5] The WAI-SR measure of therapeutic alliance and provider-patient relationship will be referred to as therapeutic relationship throughout the paper.

Survey participants were reminded that surveys were not being reviewed live, therefore they were encouraged to seek help from their physician or emergency services if they were in need of urgent mental health resources.

Statistical Analysis: Baseline characteristics were compared between the two groups using Fisher's exact test. The means and standard deviations of each of the mental health scores (EPDS, PHQ-9, GAD-7) at baseline and 6-weeks, as well as the change score from baseline to 6-weeks, were calculated. Welch's t-test was used to compare the change in the scores between the two groups, after plotting histograms to confirm normality of the score distributions. Statistical significance was set at two-sided p-value of 0.05. Our primary outcome was depression as measured by the change in PHQ-9 and EPDS from baseline to 6-weeks postpartum. To explore if women who screened positive for baseline levels of depression and anxiety responded differently between the two groups, post-hoc exploration was conducted via significance tests.

Sample size calculation: Based on power calculations, with a two-sided rejection region, alpha level of 0.05, power of 80%, projected effect size of Cohen's d=0.40, and 20% attrition, the target sample size was N=65 per group or 130 participants. The effect size of d=0.40 was selected based on a prior study of the Woebot intervention in young adults using the PHQ-9.13

The present study ultimately enrolled more participants than projected given high interest from patients for study participation.

RESULTS

From January to May 2019, 467 inpatient women were assessed for eligibility within 72 hours of delivery. Of the 282 women who were eligible, 193 (68%) consented to participate. Among this group, 192 women completed the baseline assessment and were randomized to the usual care or chatbot group (n=96 in each condition; Figure 1).

There were no differences between the two groups at baseline on any variable except for the distribution of age but mean age in both groups was 34 years (Table 1). 42.7% of the sample self-identified as White, nearly half were from minority groups (Asian/Pacific Islander=38.8%; Latinx=6%; Black=1.6%), and most were employed (77%) and married (88%). Notably, nearly 40% of the sample had previously been in therapy (37% usual care and 38.2% chatbot). Five participants reported current psychiatric medications use at baseline. In terms of opinions regarding psychiatric care (Table 1), less than half of the total sample (48.6%) were satisfied with the provision of usual care for postpartum depression and anxiety. Over 90% reported that depression and anxiety are important to monitor during pregnancy and postpartum yet nearly 60% indicated that child-bearing women face stigma if they seek depression or anxiety services. Nearly 80% endorsed being comfortable with using a mobile application for mood management.

Baseline mean scores on all clinical measures (PHQ, EPDS, and GAD) were below their respective screening thresholds for the total sample and in both groups (Table 2). Specifically, the mean EPDS score for both groups was <10 (usual care=5.37 (SD=4.20); chatbot=5.51 (SD=4.70)), indicating an unlikely presence of depression according to scoring conventions. Although there is no universal threshold, an EPDS score of 10 is a common screening positive threshold.1 Mean PHQ-9 and GAD-7 scores for both groups were <5 , (PHQ: usual care=3.36 (SD=3.05); chatbot=4.41 (SD=4.29); GAD: usual care=3.60 (SD=3.71) and chatbot=4.04 (SD=3.41)), indicating no presence of depression or anxiety according to scoring conventions.[13]

Mental Health Outcomes at 6-weeks End of Treatment

Regarding primary outcomes, there was a statistically significant difference between the two groups in mean change scores from baseline to 6-weeks for PHQ-9 (chatbot mean difference= -1.32 (SD=3.4); usual care mean difference= -0.13 (SD=3.01); p-value = 0.025; see Table 2); mean scores remained below screening thresholds at 6-weeks. There were no statistically or clinically significant differences between the two groups on the EPDS at 6-weeks and mean scores for the EPDS remained below screening thresholds.

Regarding anxiety, there were no differences between the two groups on anxiety as measured by the GAD-7 mental health assessment (Table 2) and mean GAD-7 scores were below screening

thresholds at baseline and 6-weeks. There were no significant differences on depression outcomes at 6-weeks between groups when analyses were stratified age, prior psychiatric diagnosis, current and/or past therapy, and self-reported frequency of app usage.

Subgroup Analysis of Depression

A post hoc exploration was conducted to explore if women who screened positive for baseline levels of depression responded differently between the two groups. At baseline, 55 total women had elevated PHQ-9 scores ≥ 5 which indicates the presence of at least mild symptoms of depression. Among this group, mean PHQ-9 scores decreased at 6 weeks in the chatbot group (n=28; baseline=8.18 (SD=4.29) to 6-weeks=4.93 (SD=3.27)) and were stable in the usual care group (n=27; baseline=7.07 (SD=2.35) to 6-weeks=6.52 (SD=4.61)) yielding a significant between-group difference (p-value=0.027). At baseline, 30 total women had elevated EPDS scores ≥10 which suggests the presence of at least mild symptoms of depression. Among this group, mean EPDS scores reduced at 6 weeks in the chatbot group (n=14; baseline=12.43 (SD=2.53); 6-weeks=8.64 (SD=5.45)) and the usual care group (n=16; baseline=11.94 (SD=2.17); 6-weeks=9.44 (SD=5.91)), yielding a non-significant between group difference (p-value=0.55).

Chatbot Group: Satisfaction, Application Usage, and Therapeutic Relationship Mean

CSQ-8 satisfaction score in the chatbot group was 24.0 (SD=5.7). 64% of chatbot users had scores ≥ 24, indicating high satisfaction with the tool (see Table 3). Most patients (74%) in the

chatbot group reported using the chatbot at least once in the two weeks prior to the 6-week survey completion. There were no significant differences in satisfaction or therapeutic alliance scores between participants by frequency of use (Table 3).

Therapeutic relationship as measured by the WAI-SR demonstrated high scores in the bond-subscale (mean=3.3 (SD=1.6)), indicating the participants endorsed establishing an affective bond with the chatbot. Chatbot participants had a medium score on the goal and task subscales (mean=2.9 (SD=1.5); mean=2.7 (SD=1.3) respectively).

COMMENT

Principal Findings

A general postpartum population felt that monitoring depression and anxiety in the postpartum period was important and demonstrated an interest in using a smartphone application for mood management. Moreover, the use of a chatbot delivering CBT and IPT therapeutic techniques and perinatal psychoeducational content was acceptable to women in the early postpartum period. Relative to a usual care control group, the use of a chatbot improved mean depression scores between birth and 6-weeks postpartum based on one common depression screening tool, PHQ-9, even though the sample was below the screening threshold for depression at baseline. There were no significant differences between groups on the EPDS or an anxiety

screening tool, GAD-7, although both means were below screening thresholds at baseline, indicating there was no clinical need for improvement of scores.

Comparison with Existing Literature / Results in the Context of What is Known

Mental health guided self-help programs delivered through smartphone applications offer improved ease of access and immediate scale potential, but their utility and efficacy must be evaluated empirically. Millions of women suffer from significant mood changes during the peripartum period yet only a small fraction access help. Reducing maternal depression is associated with positive outcomes for both mother and child.[14] Psychotherapy is considered first line treatment and does not involve the risk of exposure to medications,[15] yet barriers to psychotherapy are persistent and pervasive (e.g., shortage of licensed mental health providers, stigma, transportation, childcare, cost prohibitive). Thus, most women go unrecognized and under- or completely untreated. This study was the first to investigate the preliminary efficacy and acceptability of a mental health chatbot that delivers evidence-based psychotherapeutic techniques in the provision of postpartum mental health care.

This study population was on average asymptomatic for depression. Stated differently, their mean scores on two gold-standard clinical screening tools for depression were less than established cut-offs that indicate the presence of depression (PHQ-9 ≤5and EPDS ≤10). The study was not powered to detect change in a non-clinical sample. Moreover, statistically

significant reductions in depression are not expected at all in a non-depressed population, such as the present sample.

Despite this, results demonstrated that the chatbot group yielded a decrease in PHQ-9 scores at 6-weeks postpartum compared to participants randomized to the usual care group. Similarly, among women who screened positive for at least mild depression symptoms at baseline on the PHQ-9 (PHQ-9 >5; n=55, usual care n=27, chatbot n=28), there was a greater reduction in PHQ-9 scores in the chatbot group compared to the usual care at 6 weeks. There were no differences between the chatbot group and the usual care group on the EPDS among the entire sample or among those who screened positive for at least mild symptoms at baseline on the EPDS (EPDS >/=10). A likely explanation for this lack of significant between group differences in this elevated sub-group is that there was only n=30 women in this analysis (usual care n=16; chatbot n=14) which is underpowered to detect an effect. Although statistically significant differences in depression, are not expected among an asymptomatic population, the direct of improvement of scores for the chatbot over the usual care group is consistent across both the PHQ-9 and EPDS measures and analyses. A study of individuals with depression at baseline, indicated by either or both a screening tool (PHQ-9; EPDS) or a structured diagnostic clinical interview (SCID), would further inform the chatbot's true impact upon women with depression. Such a study could also utilize recommended thresholds for minimal clinically important difference (MCID) in outcomes given the presence of mild, moderate, and /or severe depression in the sample population at baseline. [23-25] Results also demonstrated that patients found the chatbot highly acceptable. Comments from participants showed the chatbot helped them to consciously focus on mood

changes and provided them with tools and techniques to impact their mood and behavior. Participants were satisfied with the chatbot due to increased availability and the ability to check in on their mental health at their own convenience at any time of the day or night. Such acceptability data among a population in need of mental health care and facing significant mood changes is promising to both patients and providers who often find themselves with limited options for mental health care referral resources. This data is consistent with a recent study which demonstrated that using text-messaging technology to screen women for postpartum depression and provide information on postpartum mental health appears to be sensitive, feasible, and well accepted.[16] The study results indicated that postpartum patients are engaged with message-based communication, with nearly all participants responding to at least one of the six text messages, and two-thirds responding to all six messages. Participants randomized to the chatbot group also established a therapeutic alliance with chatbot. Specifically, all WAI-SR scores were in the acceptable range, with specific questions about the bond between Woebot and the participant showing the highest scores. These results mirror those previously found among a large cohort of over 30,000 Woebot users[20] as well as among two distinct populations of individuals with substance abuse.[17,19] Therapeutic alliance as a whole contributes to positive psychotherapy outcomes, but this is the first study to demonstrate positive therapeutic relationship between a chatbot and postpartum women. Usage of the chatbot varied significantly among women. We did not find factors clearly associated with high usage, and depression scores did not seem to be inversely related to usage, although the study was not powered to detect differences for groups based on usage.

Strengths and Limitations

Study strengths include the randomized design and low dropout rate. Feasibility and study interest were high; women who had given birth within the previous 72 hours were approached during their birth hospitalization and offered the opportunity to participate in this study. Despite a myriad of adjustments occurring within these early postpartum days, this study yielded a 68% consent to participate rate among women who were eligible to participate. Another study strength is that nearly half of the sample self-identified as a minority population, highlighting the diversity of the group. The depression outcomes were derived from internationally recognized gold-standard self-report measures with substantial validity and reliability, and represent the most common depression screeners presently used in clinical practice and settings, unlike lengthy diagnostic measures which are often impractical to implement in real world clinical settings.

The current study has several limitations. The study period was limited to short-term outcomes with no follow-up period and inability to confirm usage of chatbot beyond self-reported usage. The study did not assess gender identity or specific relevant obstetric history factors (i.e., pregnancy and delivery complications, route of delivery) that may influence depression and anxiety symptoms; future research is encouraged to be inclusive of these demographic elements. The mean baseline scores on all the mental health screening tools were below screening thresholds, indicating an asymptotic sample and limiting the ability to detect significant decreases. Our population was more likely to be married and more educated than the general US population, and included a higher proportion of Asian American-pacific Islander participants and a lower proportion of black participants. The population that was invited to

participate in this study were women who gave birth at Lucille Packard's Children's Hospital from January 2019 to May 2019 and reflect the maternity population typically served at this hospital. Further investigation of the chatbot's feasibly, acceptability and preliminary efficacy among additional national samples is warranted. Future research on the chatbot should also sample women further out from birth, as mood concerns may be more likely to arise thereafter, as well continue to use a randomized design, perhaps with an active comparator and include a follow-up period.

Conclusions and Implications

Millions of women suffer from significant mood changes during the peripartum period yet only a small fraction access help, leaving them often alone to face the tremendously daunting and simultaneous lifestyle changes that motherhood demands. Improved access to maternal mental health care is a fundamental need that women, their families, providers, and clinics face together. Yet smartphone ownership in women of child-bearing age is very high as in interest in digital mood management options. A smartphone application-based conversational agent (chatbot) that delivers mood management programs is poised to dramatically improve access to mental health treatment and deliver support to a large number of women immediately. This randomized controlled trial demonstrated that the chatbot had high feasibility and acceptability as well as preliminary efficacy in reducing depression score on one gold-standard depression screening tool. Future studies investigating this technology within a cohort with clinical levels of

depression and over a longer period of time are encouraged, with particular attention made to addressing strengths and limitations identified in this preliminary/pilot study.

References

1. Davey HL, Tough SC, Adair CE, Benzies KM. Risk factors for sub-clinical and major postpartum depression among a community cohort of Canadian women. Matern Child Health J. 2011;15(7):866-875. doi:10.1007/s10995-008-0314-8

2. ACOG Committee Opinion No. 757: Screening for Perinatal Depression. Obstet Gynecol. 2018;132(5):e208-e212. doi:10.1097/AOG.0000000000002927

3. Gelenberg AJ, Freeman MP, Markowitz JC, et al. APA Practice Guidelines for the Treatment of Patients with Major Depressive Disorder. Published online 2010:152.

4. Cuijpers P, Brännmark JG, van Straten A. Psychological treatment of postpartum depression: a meta-analysis. J Clin Psychol. 2008;64(1):103-118. doi:10.1002/jclp.20432

5. Claridge AM. Efficacy of systemically oriented psychotherapies in the treatment of perinatal depression: a meta-analysis. Arch Womens Ment Health. 2014;17(1):3-15. doi:10.1007/s00737-013-0391-6

6. Sockol LE. A systematic review of the efficacy of cognitive behavioral therapy for treating and preventing perinatal depression. J Affect Disord. 2015;177:7-21. doi:10.1016/j.jad.2015.01.052

7. Stuart S. Interpersonal psychotherapy for postpartum depression. Clin Psychol Psychother. 2012;19(2):134-140. doi:10.1002/cpp.17781 Guidance | Depression in adults: recognition and management | Guidance | NICE. Accessed July 20, 2020. https://www.nice.org.uk/guidance/cg90/chapter/1-Guidance

8. Gaffney H, Mansell W, Tai S. Conversational Agents in the Treatment of Mental Health Problems: Mixed-Method Systematic Review. JMIR Ment Health. 2019;6(10). doi:10.2196/14166 NW 1615 L. St, Suite 800 Washington, Inquiries D 20036USA202-419-4300 | M-857-8562

9. F-419-4372 | M. Demographics of Mobile Device Ownership and Adoption in the United States. Pew Research Center: Internet, Science & Tech. Accessed June 9, 2021. https://www.pewresearch.org/internet/fact-sheet/mobile/

10. Fitzpatrick KK, Darcy A, Vierhile M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. JMIR Ment Health. 2017;4(2):e19. doi:10.2196/mental.7785

11. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340 (mar23 1):c869-c869. doi:10.1136/bmj.c869

12. Instructions for Patient Health Questionnaire (PHQ) and GAD-7 Measures. https://www.ons.org/sites/default/files/PHQandGAD7_InstructionManual.pdf

13. Shaw DS, Connell A, Dishion TJ, Wilson MN, Gardner F. Improvements in maternal depression as a mediator of intervention effects on early childhood problem behavior. Dev Psychopathol. 2009;21(2):417-439. doi:10.1017/S0954579409000236

14. Fitelson E, Kim S, Baker AS, Leight K. Treatment of postpartum depression: clinical, psychological and pharmacological options. Int J Womens Health. 2010;3:1-14. Published 2010 Dec 30. doi:10.2147/IJWH.S6938

15. Lawson A, Dalfen A, Murphy KE, Milligan N, Lancee W. Use of Text Messaging for Postpartum Depression Screening and Information Provision. Psychiatr Serv Wash DC. 2019;70(5):389-395. doi:10.1176/appi.ps.201800269

16. Prochaska JJ, Vogel EA, Chieng A, et al. A Therapeutic Relational Agent for Reducing Problematic Substance Use (Woebot): Development and Usability Study. J Med Internet Res. 2021;23(3):e24850. Published 2021 Mar 23. doi:10.2196/24850

17. Robinson A, Eaneff S., and Darcy A. Presentation 3: RCT of Woebot for Adolescent Depression compared to Digital Psychoeducation: The Headway Study. Diversity in Digital

Mental Health Interventions. Association of Behavioral and Cognitive Therapies, New York, November 2022.

18. Prochaska JJ, Vogel EA, Chieng A, et al. A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. Drug Alcohol Depend. 2021;227:108986. doi:10.1016/j.drugalcdep.2021.108986

19. Darcy A, Daniels J, Salinger D, Wicks P, Robinson A. Evidence of Human-Level Bonds Established with a Digital Conversational Agent: Cross-sectional, Retrospective Observational Study. JMIR Form Res. 2021 May 11;5(5):e27868. doi: 10.2196/27868.

20. Darcy A, Beaudette A, Chiauzzi E, Daniels J, Goodwin K, Mariano TY, Wicks P, Robinson A. Anatomy of a Woebot® (WB001): agent guided CBT for women with postpartum depression. Expert Rev Med Devices. 2022 Apr;19(4):287-301. doi: 10.1080/17434440.2022.2075726. PMID: 35748029.

21. Wang L, Kroenke K, Stump TE, Monahan PO. Screening for perinatal depression with the Patient Health Questionnaire depression scale (PHQ-9): A systematic review and meta-analysis. Gen Hosp Psychiatry. 2021 Jan-Feb;68:74-82. doi: 10.1016/j.genhosppsych.2020.12.007. Epub 2020 Dec 21. PMID: 33360526; PMCID: PMC9112666.

22. Sit DK, Wisner KL. Identification of postpartum depression. Clin Obstet Gynecol. 2009 Sep;52(3):456-68. doi: 10.1097/GRF.0b013e3181b5a57c. PMID: 19661761; PMCID: PMC2736559.

23. Mao F, Sun Y, Wang J, Huang Y, Lu Y, Cao F. Sensitivity to change and minimal clinically important difference of Edinburgh postnatal depression scale. Asian J Psychiatr. 2021 Dec;66:102873. doi: 10.1016/j.ajp.2021.102873. Epub 2021 Sep 29. PMID: 34624746

24. Matthey, Stephen. (2004). Calculating clinically significant change in postnatal depression studies using the Edinburgh Postnatal Depression Scale. Journal of affective disorders. 78. 269-72. 10.1016/S0165-0327(02)00313-0.

25. Löwe B, Kroenke K, Herzog W, Gräfe K. Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9) Journal of Affective Disorders. 2004 Jul;81(1):61–66. doi: 10.1016/S0165-0327(03)00198-8.

Figure 1. Patient flow throughout the study enrollment and assessment processes

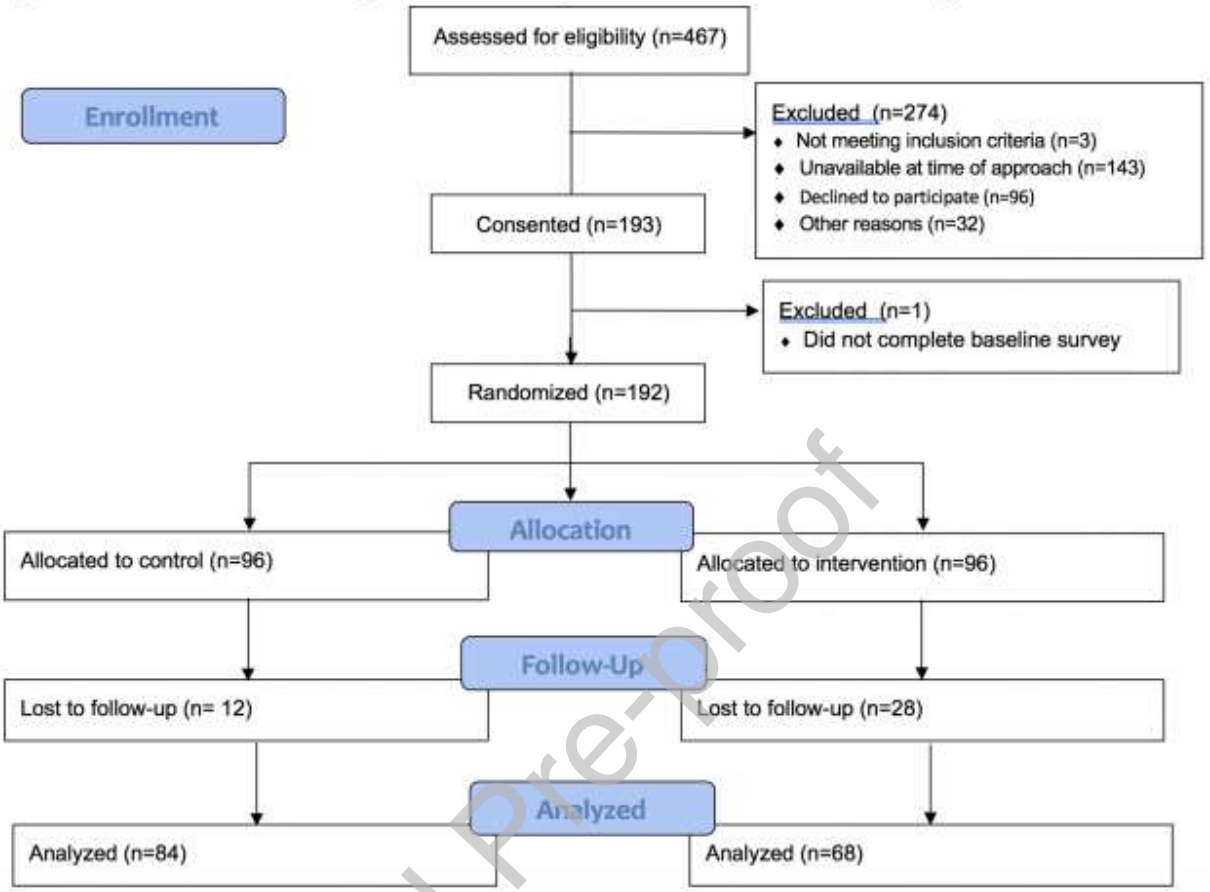
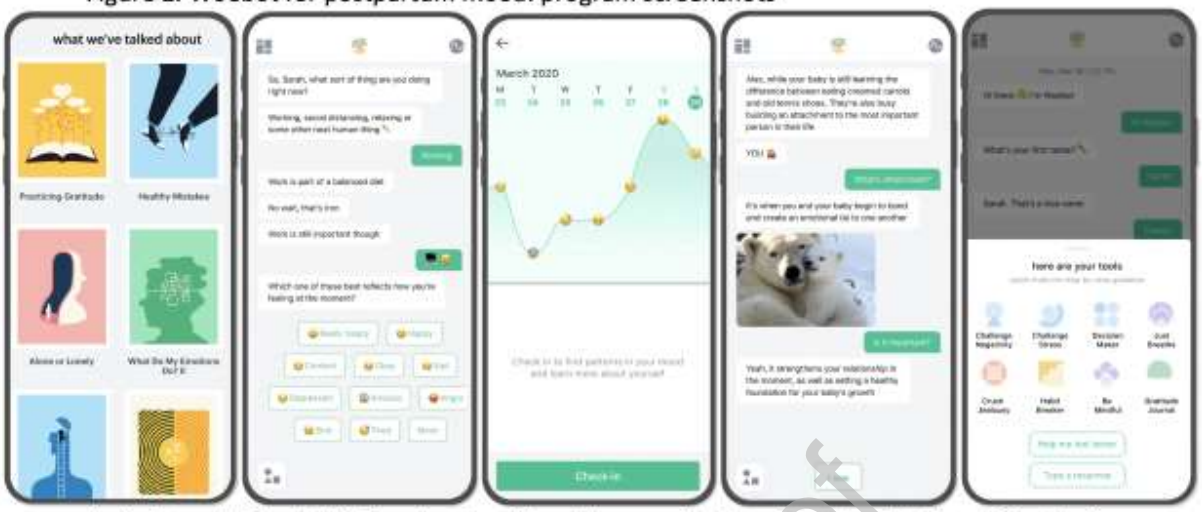


Figure 2. Woebot for postpartum mood: program screenshots



Note. From left to right i. Psychoeducational Lessons; ii. Mood Tracking; iii. Mood Graph of previously entered moods; iv. A text-based conversation in the application; v. cognitive behavioral therapy and interpersonal psychotherapy tools for mood management.

Table 1. Baseline demographics, psychiatric history, and opinions regarding psychiatric care by group

Variable	Chatbot Group (n=68) % (n)	Usual Care Group (n=84) % (n)	p-value
Age in Years			0.02
18-24	0 (0)	7.1% (6)	
25-34	63.2% (43)	47.6% (40)	
35-44	32.4% (22)	44.0% (37)	
45+	4.4% (3)	1.2% (1)	
Race-Ethnicity			0.14
White	48.5% (32)	38.1% (32)	
Asian/Pacific Islander	39.7% (27)	38.1% (32)	
Black/African American	1.6% (1)	1.2% (1)	
Hispanic/Latinx	6% (4)	4.8% (4)	
Other	5.9% (4)	19.0% (15)	
Martial Status			0.046
Married	94.1% (64)	83.3% (70)	
Not Married	5.9% (4)	16.7% (14)	
Employment Status			
Employed Part or Full-Time	80.9% (55)	78.6% (66)	0.14
Unemployed (including homemaker, student, retired)	19.1% (13)	21.4% (18)	0.84
Psychiatric History			
Pre-existing mental health condition	16.2% (11)	9.5% (8)	0.23
Psychiatric medication use	5.9% (4)	1.2% (1)	0.87
Previous use of psychotherapy	38.2% (26)	36.9% (31)	0.87
Current use of psychotherapy	30.8% (21)	19.4% (17)	0.37
Opinions Regarding Psychiatric Care at Enrollment			
Current use of a smartphone app to support mental health	4.4% (3)	10.7% (9)	0.55
Comfortable with use of mobile app to support mental health	82.4% (56)	77.4% (65)	0.74
Satisfied with medical services in pregnancy specifically regarding anxiety and depression	47.1% (32)	50.0% (42)	0.89

Anxiety and depression are an important part of a person's general health	88.2% (60)	78.6% (66)	0.13
Anxiety and depression are important to monitor during pregnancy and postpartum	97.1% (66)	88.1% (74)	0.07
Child-bearing women face stigma if they seek depression or anxiety services	58.5% (40)	59.5% (50)	0.97

Note. Mean age in each group was 34 years. 100% of unmarried were never married in chatbot and 73% in usual care (3 divorced or separated).

Table 2. Mental health outcomes by group

	Chatbot Group (n=68)			Usual Care Group (n=84)			
Variable	**BL M(SD)**	**6-weeks M(SD)**	**Change Score**	**BL M(SD)**	**6-weeks M(SD)**	**Change Score**	**p-value**
Entire Sample							
PHQ-9	4.41 (4.29)	3.09 (3.02)	-1.32 (3.40)	3.36 (3.05)	3.23 (3.84)	-0.13 (3.01)	0.025*
EPDS	5.51 (4.70)	4.88 (5.26)	-0.63 (5.29)	5.37 (4.20)	4.61 (5.20)	-0.76 (4.11)	0.87
GAD-7	4.04 (3.41)	3.32 (3.85)	-0.72 (3.35)	3.60 (3.71)	3.63 (4.45)	.036 (3.78)	0.19
Elevated Sample							
PHQ-9 (≥5)	8.18 (4.29); n=28	4.93 (3.27)	-3.25 (4.17)	7.07 (2.35); n=27	6.52 (4.61)	-0.55(4.60)	0.027*
EPDS (≥10)	12.43 (2.53); n=14	8.64 (5.45)	-3.79 (6.33)	11.94 (2.17); n=16	9.44 (5.91)	-2.50 (5.03)	0.55

Note. BL=baseline; PHQ-9=Patient Health Questionnaire-9; scores of 0-4=minimal/no depression, 5-9=mild depression, 10-14=moderate depression, 15+=moderately-severe or severe depression; EPDS=Edinburgh Postnatal Depression Scale; although there is no universal threshold, an EPDS score of 10 is a common screening positive threshold for possible depression; GAD=Generalized Anxiety Disorder-7; the GAD-7 is scored identically to PHQ-9, 0-4=minimal or no anxiety, 5-9=mild anxiety, 10-14=moderate anxiety, and 15+=moderately-severe or severe anxiety.

Table 3. Treatment Participants' Satisfaction and Alliance Ratings for Woebot (n =67)

	Possible Range	Mean (SD)
Client Satisfaction Questionnaire (CSQ-8, 8 items)	**1-4**	
How would you rate the quality of service you have received?		3 (0.83)
Did you get the kind of service you wanted?		2.73 (0.75)
To what extent has our program met your needs?		2.53 (0.92)
If a friend were in need of similar help, would you recommend our program to him or her?		2.88 (0.81)
How satisfied are you with the amount of help you have received?		2.76 (0.78)
Have the services you received helped you to deal more effectively with your problems?		2.77 (0.73)
In an overall, general sense, how satisfied are you with the service you have received?		2.74 (0.84)
If you were to seek help again, would you come back to our program?		
Total score		2.85 (0.87)
	8-32	22.3 (5.7)
Working Alliance Inventory-Short Revised (WAI-SR)		
Agreement with treatment goals (4 items)	1-5	2.93 (1.49)
Agreement with treatment tasks (4 items)	1-5	2.66 (1.31))
Affective bond formation in treatment (4 items)	1-5	3.33 (1.57)
Total score	1-5	2.97 (1.35)

Note. Higher numbers indicate more higher satisfaction and higher alliance. One participant did not respond to the satisfaction and alliance questionnaires.