Artificial Intelligence in Depression–Medication Enhancement (AID-ME):

A Cluster Randomized Trial of a Deep-Learning-Enabled Clinical Decision Support System for Personalized Depression Treatment Selection and Management

David Benrimoh, MD, CM, MSc, FRCPC; Kate Whitmore, MSc; Maud Richard, MSc, RAC; Grace Golden, MSc; Kelly Perlman, PhD; Sara Jalali, MSc; Timothy Friesen, MSc; Youcef Barkat, MSc, MSc; Joseph Mehltretter, MSc; Robert Fratila; Caitrin Armstrong, MSc; Sonia Israel; Christina Popescu, MSc; Jordan F. Karp, MD; Sagar V. Parikh, MD, FRCPC; Shirin Golchi, MSc, PhD; Erica E. M. Moodie, MPhil, MSc, PhD; Junwei Shen, MSc, PhD; Anthony J. Gifuni, MD, MSc, FRCPC; Manuela Ferrari, PhD; Mamta Sapra, MD; Stefan Kloiber, MD; Georges-F. Pinard, MD, FRCPC; Boadie W. Dunlop, MSc, MD; Karl Looper, MD, MSc; Mohini Ranganathan, MD; Martin Enault; Serge Beaulieu, MD, PhD, FRCPC; Soham Rej, MD, MSc, FRCPC; Fanny Hersson-Edery, MD; Warren Steiner, MD, FRCPC, FAPA; Alexandra Anacleto; Sabrina Qassim; Rebecca McGuire-Snieckus, MSc, PhD, FRSS; and Howard C. Margolese, MSc, MD, CM, FRCPC

Abstract

Background: There has been increasing interest in the use of artificial intelligence (AI)-enabled clinical decision support systems (CDSS) for the personalization of major depressive disorder (MDD) treatment selection and management, but clinical studies are lacking. We tested whether a CDSS that combines an AI which predicts remission probabilities for individual antidepressants and a clinical algorithm based on treatment can improve MDD outcomes.

Methods: This was a multicenter, cluster randomized, patient-and-rater blinded and clinician-partially-blinded, active-controlled trial that recruited outpatient adults with moderate or greater severity MDD. All patients had access to a patient portal to complete questionnaires. Clinicians in the active group had access to the CDSS; clinicians in the active-control group received patient questionnaires; both groups received guideline training. Primary outcome was remission (<11 points on the Montgomery-Asberg Depression Rating Scale [MADRS]) at study exit.

Results: Forty-seven clinicians were recruited at 9 sites. Of 74 eligible patients, 61 patients completed a postbaseline MADRS and were analyzed. There were no differences in baseline MADRS ($P = .153$). There were more remitters in the active (n = 12, 28.6%) than in the active-control (0%) group ($P = .012$, Fisher's exact). Of 3 serious adverse events, none were caused by the CDSS. Speed of improvement was higher in the active than the control group (1.26 vs 0.37, $P = .03$).

Conclusions: While limited by sample size and the lack of primary care clinicians, these results demonstrate preliminary evidence that longitudinal use of an AI-CDSS can improve outcomes in moderate and greater severity MDD.

Trial Registration: ClinicalTrials.gov identifier: NCT04655924

J Clin Psychiatry 2025;86(3):24m15634

Author affiliations are listed at the end of this article.

Major depressive disorder (MDD) is a leading cause of disability and socioeconomic burden[1] impacting more than 300 million people worldwide.[2] Unfortunately, only a minority of patients will improve with the first treatment trial, and repeated treatment trials have diminishing probabilities of success.[3] Many patients undergo an arduous "trial and error" treatment selection approach, resulting in poorer outcomes, longer time in treatment, and greater patient and family burden.[4] To improve outcomes, it would be

Scan Now

See supplementary material for this article at Psychiatrist.com

Editor's Note

We encourage authors to submit papers for consideration as a part of our Early Career Psychiatrists section. Please contact Joseph F. Goldberg, MD, at psychiatrist.com/contact/goldberg.

Clinical Points

- There has been increasing interest in using artificial intelligence (AI) to assist mental health clinical decisions, but few trials of patient outcomes.
- In this trial, we demonstrated that a clinical decision support system combining AI treatment predictions and best practices improved remission for depressed outpatients.
- Patients in the active group improved faster; no serious adverse events were attributed to the intervention.

valuable to have a scalable point-of-care tool that can help personalize treatment choice without requiring expensive testing.[5,6]

There have been several efforts in recent years to use artificial intelligence (AI) to predict treatment outcomes in order to better match patients to specific treatments (see Squarcina et al[7]). Most studies have differentiated between 2 treatments (eg, 2 drugs) or treatment types (eg, 2 types of psychotherapy), limiting clinical utility when many treatments are available. In addition, clinicians are often concerned about model bias and being able to interpret the outputs of AI predictive models which are often considered to be "black boxes"[8–10] (providing predictions without clear reasoning). Furthermore, while improving predictions about treatment outcome may be helpful to personalize treatment, previous work has shown that treatments are often not managed in accordance with guidelines in terms of dosage and monitoring.[11–13] There is a need for a solution to both the treatment *selection* and the treatment *management* problems while integrating into existing clinical workflows.[14]

To address this, Aifred investigators developed the Aifred clinical decision support system (hereinafter referred to as the CDSS). This is a digital platform that supports clinicians in the implementation of guidelines (2016 Canadian Network for Mood and Anxiety Treatments [CANMAT] depression guidelines[15]) and measurement-based care[15] in order to solve the treatment *management* problem, and which includes an AI (deep learning)-powered module to assist in baseline treatment *selection* by providing predicted probabilities of remission for 10 commonly used first-line antidepressants and combinations of these. Extensive feasibility and ease of use testing of this CDSS was previously performed in both simulation center and in vivo feasibility studies.[16–19] With in silico testing demonstrating that the AI component should help improve remission rates[6,20–22] and in vivo testing demonstrating that the platform was feasible, easy to use, and likely safe,[16–19] the current study was undertaken with the main objective of determining the efficacy of the platform in improving depression treatment outcomes in patients with moderate to severe depression, as well as to assess platform safety.

METHODS

This study is reported as per the CONSORT-AI checklist[23] and was conducted in accordance with all relevant ethical regulations including the Declaration of Helsinki and the Tri-Council Policy Statement. The research ethics board of the Douglas Research Center gave ethical approval for this work, and it was subsequently approved by central and local ethics review boards for each site. Written informed consent was obtained from all study participants. Full methods are available in the supplementary material (available on request from the corresponding author).

Design

The current study is a 2-arm, cluster-randomized trial, with clinicians serving as the cluster. Clinicians rather than patients were randomized as they were the ones receiving the decision support intervention, and to avoid contamination.[24] Patients entered the intervention arm of their treating clinician.

Participants: Clinicians

Clinicians could include primary care doctors, psychiatrists, residents, and nurse practitioners (with or without specialized mental health training), who saw at least 1 patient with depression per month, on average, before study start.

Participants: Patients

Patient recruitment criteria were intended to be broad in order to replicate a naturalistic outpatient depression population with moderate to severe depression. Patients were recruited from the practices or hospital-based clinics of the participating clinicians. Inclusion criteria were as follows: (1) age 18 years and over; (2) diagnosed by their treating clinician with MDD using *DSM-5* criteria[25]; (3) MDD diagnosis confirmed via a blinded rater who completed the Mini-International Neuropsychiatric Interview (MINI)[26]; and (4) at least moderate severity, as assessed by a blinded rater completing the Montgomery-Asberg Depression Rating Scale (MADRS) (using a cutoff of 20).[27] Exclusion criteria were as follows: (1) age under 18 years; (2) presence of bipolar disorder of any type; (3) inability or unwillingness to give informed consent; (4) inability to manage patient safely as an outpatient; (5) an active major depressive disorder was not the main condition being treated; and (6) inability to use the tool (eg, because of severe cognitive impairment). An active major depression meant that the depression, in the judgment of the treating clinician, required an initiation or a change in treatment. Psychiatric comorbidities (aside from bipolar disorder) were permitted.

Settings

Eligible settings included any primary or secondary/tertiary public or private outpatient setting in the United States or Canada that provided outpatient care for patients presenting with MDD. A diverse array of sites joined the study. These included public sector psychiatric clinics in Canada and university-affiliated and Veterans Affairs mental health services in the U.S.

Intervention—Aifred CDSS

The Aifred CDSS platform, the intervention in this study, consists of the following: (1) the patient portal, accessible by web browser or mobile phone application, allows patients to complete questionnaires, receive email reminders to complete questionnaires, visualize questionnaire scores and interpretations, and track treatments which they or their clinicians enter and (2) the clinician portal, accessible by web browser, allows clinicians to see all the information patients enter. This clinician portal includes several important modules.

First, it includes a measurement-based care component where clinicians can see the data that patients enter.

Second, it has a clinical algorithm module. This module is a rule-based decision tree based on the CANMAT 2016 guidelines for depression treatment.[15] This module presents the clinician with patient-specific, guideline-derived information about treatment options based on the patient's depression severity, change in depression severity over time (measured using the Patient Health Questionnaire-9 [PHQ-9][28]), and current treatments. The algorithm provides new information at each patient visit, based on patient progress, and provides standard dosing information based on the CANMAT guidelines.

Third, it has an AI component. This AI component is separate from the clinical algorithm based on the guidelines, but it is housed within the algorithm module. The AI is focused on assisting with treatment selection by generating remission probabilities for 8 commonly used first-line antidepressants (citalopram, paroxetine, duloxetine, venlafaxine, fluoxetine, bupropion, sertraline, and escitalopram) and 2 commonly used combinations of first-line antidepressants (venlafaxine-XR plus mirtazapine, and escitalopram plus bupropion). At the point where the clinical algorithm presents a page for the clinician to select treatment, these probabilities are displayed, and in accordance with clinician feedback during development, the treatments are ranked in order of their probability of remission. Remission probabilities were calculated using our model which incorporated data from 9,042 patients from 21 depression treatment studies.[29] All other treatments present in the guidelines which do not have treatment probabilities associated with them are also presented on the page; these were treatments for which no or insufficient training data were available. These remission probabilities are presented as a percentage probability that a given patient will remit if prescribed a given treatment, with a different probability calculated for each treatment. Clinicians were therefore provided with a list of treatments and their probabilities of remission but were able to choose whatever treatment they wished. Dosing information is presented by the clinical algorithm; the AI does not make predictions about dose but assumes (based on the training data used) that minimum therapeutic doses, per the guidelines, will be prescribed. The AI is a deep learning model trained on 9,042 patients from depression treatment trials, with remission as the training objective—in other words, it is a machine learning model (specifically a deep neural network) that uses clinical and demographic information about individual patients to make predictions about treatment outcome. Details on the methods used to develop the model are available in Perlman et al,[30] and a full discussion of the model used in this trial is available in Benrimoh et al,[29] including the input clinical and demographic features. All input features were provided by patients and clinicians via short questionnaires, such that the results from the predictive model could be available as soon as these questionnaires were completed. Further details on the AI, the data it intakes, as well as an example of the report it provides clinicians are provided in the supplementary material.

Intervention—Patients

All patients received access to the patient portal of the CDSS. Patients remained in the study for 12 weeks from their first treatment visit. Study visits occurred at baseline, week 2, at 1 point during weeks 4–6 (the timing of this visit was flexible to accommodate clinician and patient scheduling), at week 8, and at week 12.

Intervention—Clinicians

There were 2 intervention groups: an active group and an active-control group.[31,32] The active-control group was provided with all the tools required to perform best-practice measurement-based and guideline-informed care.[1] Clinicians in the active-control were provided with the results of questionnaires patients completed in PDF format as well as training on the guidelines. Guideline training involved a PowerPoint presentation by D.B. on the guideline document as well as provision of the clinician with a copy of the CANMAT guidelines. Active-control group clinicians were not required to use the information they were provided in any specific manner, in keeping with the naturalistic objectives of the study. Active group clinicians received guideline training and were provided with full access to the clinician portal of the Aifred CDSS (including the measurement-based care, clinical algorithm, and AI modules described above). They were provided with training on the CDSS platform as well as the same training on the guidelines as the control group by DB. Active clinicians, much like the active-control group, were also not required to use the information provided to them or to adhere to the AI's predictions or to the guideline information provided by the clinical algorithm. They were required to at least log in to the CDSS at each visit. While clinicians were not removed from the study if they failed to log in at each visit, they were reminded to do so. Further details are available in the supplementary material.

As all raw data provided to clinicians were the same in the active and active-control groups, and patients in both groups received the patient-portal portion of the intervention, the only group differences consisted of the provision of the data processed by the clinical algorithm and AI model to the active group. In addition, the active-control group was given the tools to approximate the clinical algorithm as they were trained on the guidelines and provided with regular questionnaire data. The active-control group was an "active" control because clinicians in that group were provided all the tools required to engage in gold-standard measurement-based care,[1] guided by the algorithm contained within the guidelines on which they were trained. In addition, patients in the control group had access to the patient platform, which would not have been the case in a treatment-as-usual control.

Measures

At baseline, the blinded rater assessed the MINI and MADRS. Patients were asked to complete the PHQ-9 and General Anxiety Disorder-7 (GAD-7) weekly once they had accounts on the CDSS. The MADRS was administered by the blinded rater at screening, visit 3 (weeks 4–6 of treatment), visit 4 (week 8), and visit 5 (week 12). Trained study staff also administered the Brief Adherence Rating Scale (BARS) after every visit to assess treatment adherence.[33] Further measures are detailed in the supplementary material.

Outcomes

The prespecified and sole primary outcome of the study was remission of depressive symptoms, defined as a score of <11[34,35] on the MADRS at study exit for those patients with at least 2 MADRS scores. Safety outcomes included an examination of the nature and number of adverse and serious adverse events in each group. Secondary outcomes included response (50% reduction in symptoms) on the MADRS, rate of change of the MADRS score, and medication adherence using the BARS score. Per the Statistical Analysis Plan, secondary outcomes were not corrected for multiple comparisons.

Blinding

Patients. Patients were fully blinded to group assignment. They were told that they were entering a study where they would be using a new digital technology to provide information to their clinician and that there were 2 groups, but the nature of the groups was not revealed.

Clinicians. Clinicians were aware of their group assignments as they were the ones receiving the AI predictions. Clinicians were partially blinded to reduce expectation bias[36]: this partial blinding involved clinicians not being told the study end points and not being informed of the expected effect sizes of the active or control interventions.

Raters. Raters who collected the primary outcome (MADRS) and conducted the MINI were blind to group allocation. Blinded raters were recruited from the research teams at participating sites by each site primary investigator. They were required to have at a minimum a bachelor's degree in a relevant field (eg, nursing, psychology, social work) and were trained to complete the MINI and MADRS interviews. For the MINI, they were required to complete the MINI training course offered by the Harm Research Institute (https://harmresearch.org/) and to either have previous, documented experience administering the MINI in clinical studies or to be trained by a clinician or someone experienced in the administration of the MINI and to complete a mock interview with the same diagnoses as a clinician or experienced rater. For the MADRS, they were required to either have previous documented experience administering the MADRS in clinical studies or they were required to read the MADRS structured interview guide[37]; all MADRS raters were then required to complete a mock administration of the MADRS where their severity rating agreed with that of a clinician or someone with experience administering the MADRS. In addition to this, all raters were required to use the MADRS structured interview[37] during rating.

Statistical Analysis

Outcome data were analyzed, as prespecified in the Statistical Analysis Plan, on an intent-to-treat basis for patients who had at least 2 ratings of the MADRS (the Analysis set). Safety data were analyzed for the Safety population, prespecified as all patients who attended at least the first treatment visit. Demographic and baseline clinical data are presented in Table 1. The primary outcome (MADRS remission) was assessed using a Fisher's exact test. Secondary outcomes were compared using 1-way analyses of variance (ANOVAs), and proportions were compared using 2-sided chi-square or 2-sided Fisher's exact tests, as appropriate. Further prespecified analyses are detailed in the Supplementary Material (available on request from the corresponding author).

Table 1.

Baseline Clinical and Demographic Characteristics per Group[a]

	Active group (n = 42)	Active-control group (n = 19)
Age, mean (SD), y	44.0 (15.2)	39.3 (12.4)
Sex, female, n (%)	20 (47.6)	10 (52.6)
Baseline MADRS, mean (SD)	33 (7.3)	30 (5.8)
Race	n = 40	n = 18
White	30 (75.0)	13 (72.2)
Other[b]	10 (25.0)	5 (27.8)
Yearly household income	n = 37	n = 15
Mean (SD), USD	42,206.87 (29,103.3)	44,817.10 (17,975.4)
Highest level of education achieved[c]	n = 40	n = 17
Some high school/high school diploma or equivalent (GED), n (%)	5 (15)	5 (27.8)
Some university or college, n (%)	12 (30.0)	4 (22.2)
Bachelor's degree, n (%)	11 (27.5)	6 (33.3)
Graduate or professional degree, n (%)	7 (17.5)	1 (5.6)
Trade/technical training or other, n (%)	5 (12.5)	1 (5.6)
Graduated high school, n (%)	39 (97.5)	17 (94.4)
Currently employed	n = 39	n = 18
n (%)	18 (41.5)	12 (66.7)
Marital status, n (%)		
Single	22 (55.0)	12 (66.7)
Partnered	18 (45.0)	6 (33.3)
No. of medications (all indications) prescribed at baseline (n = 57), mean	3.18	3.06
Adverse childhood experiences, mean (SD)	2.62 (2.59)	3.61 (2.59)
MINI comorbidities, n (%)		
Suicidality (current-past month)	22 (53.5)	13 (68.4)
High suicidality score category	15 (35.7)	6 (31.6)
Generalized anxiety disorder current	15 (35.7)	9 (47.4)
Social anxiety disorder current	11 (26.2)	4 (21.1)
Posttraumatic stress disorder current	10 (23.8)	1 (5.3)
Panic disorder current	7 (16.7)	2 (10.5)
Alcohol use disorder past 12 mo	6 (14.3)	6 (31.6)
Agoraphobia current	5 (12)	2 (10.5)
Substance use disorder (nonalcohol), past 12 mo	5 (11.9)	2 (10.5)
SAPAS-SA (personality disorder screening), n (%)	n = 41	n = 17
Those meeting cutoff score of 3 or more for positive screening	25 (61)	8 (47.1)

[a] Further demographic details are available in the Supplementary Materials (available on request from the corresponding author).

[b] Lower count rows have been collapsed into the "other" category in order to preserve confidentiality.

[c] Participants could select more than 1 option; the graduated high school entry was constructed based on the available data.

Abbreviations: GED = General Educational Development, MADRS = Montgomery-Asberg Depression Rating Scale, SAPAS-SA = Standardised Assessment of Personality Abbreviated Scale–Self Assessment, USD = United States Dollar.

Early Study Termination

Unfortunately, due to lack of funding caused by delays related to the COVID-19 pandemic, the study was terminated early.

RESULTS

Sites

Sites were located in Canada (5) and the United States (4) and included US Veterans Affairs hospitals and mood disorders programs in university-affiliated psychiatric departments.

Recruitment: Clinicians

Of the 47 clinicians recruited who were cleared to recruit patients prior to early study termination, 25 were randomized to the active group and 22 to the active-control group. 27 clinicians recruited at least 1 patient (57%); 16 in the active group (64%) and 11 in the active-control group (50%).

Recruitment and Dropout: Patients

Patients were recruited between June 15, 2022, and November 16, 2023, a total of 17 months. The study was terminated early because of lack of funds due to delays in study initiation related to COVID-19. Recruitment and dropout are summarized in the CONSORT diagram (Figure 1). Of the 74 eligible patients after screening and enrollment, 61 had at least 2 MADRS available, forming the Analysis set (n = 42 active, n = 19 active-control). The groups did not differ in terms of 12-week completion (active = 36/53 [68%]; active-control = 18/21 [86%]) (*P* = .15, Fisher's exact test).

Figure 1.
Patient CONSORT Diagram

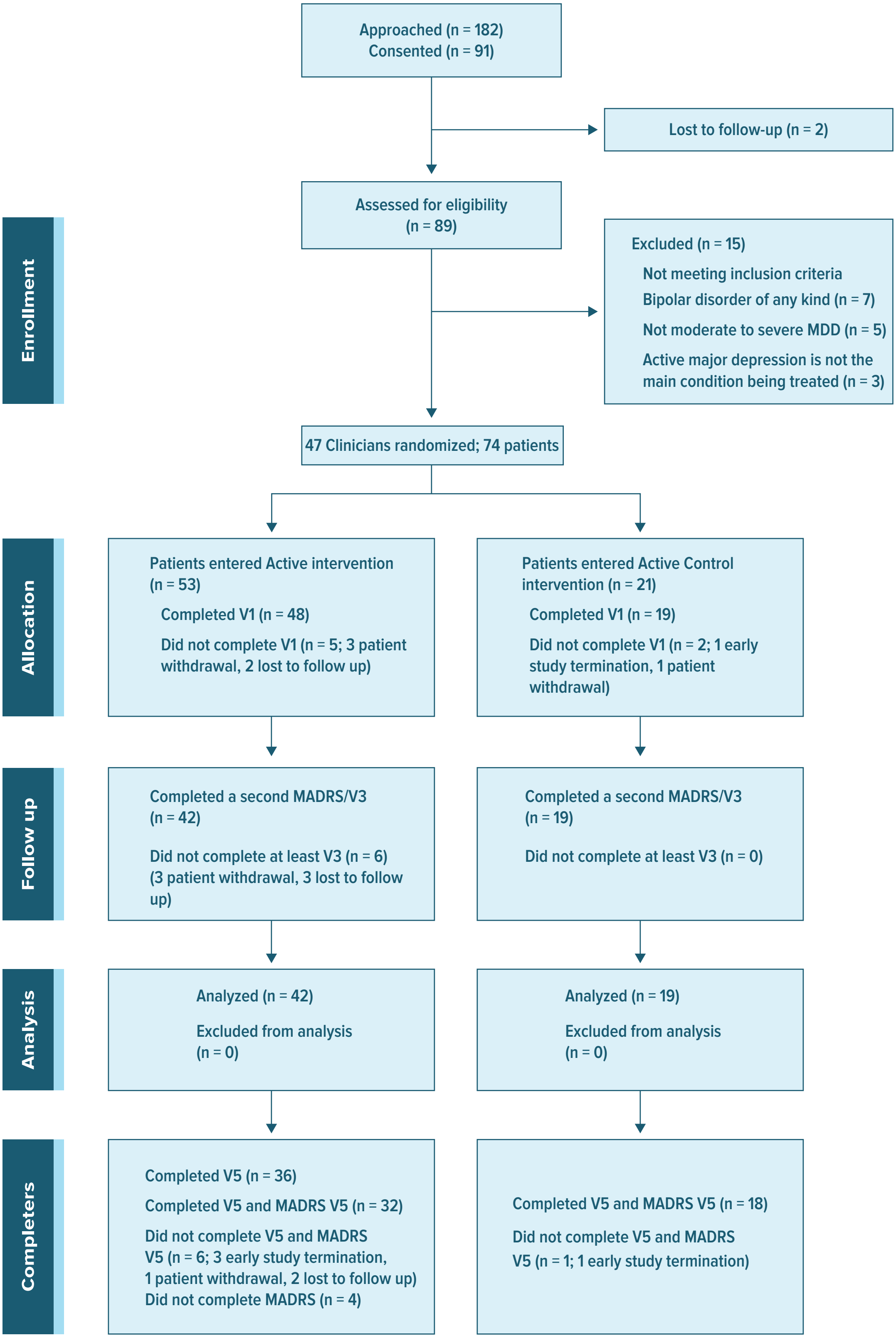

Approached (n = 182)
Consented (n = 91)
Lost to follow-up (n = 2)
Assessed for eligibility (n = 89)
Excluded (n = 15)
Not meeting inclusion criteria
Bipolar disorder of any kind (n = 7)
Not moderate to severe MDD (n = 5)
Active major depression is not the main condition being treated (n = 3)
47 Clinicians randomized; 74 patients
Enrollment
Allocation
Patients entered Active intervention (n = 53)
Completed V1 (n = 48)
Did not complete V1 (n = 5; 3 patient withdrawal, 2 lost to follow up)
Patients entered Active Control intervention (n = 21)
Completed V1 (n = 19)
Did not complete V1 (n = 2; 1 early study termination, 1 patient withdrawal)
Follow up
Completed a second MADRS/V3 (n = 42)
Did not complete at least V3 (n = 6) (3 patient withdrawal, 3 lost to follow up)
Completed a second MADRS/V3 (n = 19)
Did not complete at least V3 (n = 0)
Analysis
Analyzed (n = 42)
Excluded from analysis (n = 0)
Analyzed (n = 19)
Excluded from analysis (n = 0)
Completers
Completed V5 (n = 36)
Completed V5 and MADRS V5 (n = 32)
Did not complete V5 and MADRS V5 (n = 6; 3 early study termination, 1 patient withdrawal, 2 lost to follow up)
Did not complete MADRS (n = 4)
Completed V5 and MADRS V5 (n = 18)
Did not complete V5 and MADRS V5 (n = 1; 1 early study termination)

Table 2.
Summary of Outcomes

Outcome	Active	Active control	*P* value
Remission	12 (28.6%)	0 (0%)	.01
Response	17 (40.5%)	3 (15.8%)	.06
Mean change from baseline	12 (SD = 13.5)	4.9 (SD = 10.9)	.05
Percent change from baseline	35 (SD = 41.1)	13.2 (36.2)	.05
Slope of improvement (amount of change per week)	1.26 (SD = 1.63)	0.37 (SD = 0.91)	.03

Demographics and Baseline Clinical Characteristics

Intervention and active-control groups did not differ with regard to important clinical and demographic characteristics (Table 1). The group had substantial chronicity of illness with 33 active patients (78.6%) and 13 active-control patients (68.4%) having recurrent MDD. Further details are available in the Supplementary Material.

Treatment Outcome—Remission at Study Exit

On the primary outcome, remission, there were significantly more remitters in the active (n = 12/42 [28.6%]) than in the active-control (0/19 [0%]) group (*P* = .01, Fisher's exact test). Outcomes are summarized in Table 2.

Treatment Outcome—Response and Change from Baseline

With respect to treatment response (defined as a 50% or greater decrease in total MADRS score between screening and study exit), 17 active patients (40.5%) and 3 active-control patients (15.8%) responded at study exit. This was a large numerical difference; however, it did not reach significance ($\chi^2 = 3.6$, *P* = .06). The proportion of responders was not significantly different at visit 3 or 5 but was significantly different at visit 4 (*P* = .04, Fisher's exact test).

With respect to change from baseline to study exit, patients in the active group experienced a mean 12.0 point improvement in MADRS score (SD = 13.5) while those in the active-control group experienced a change of 4.9 (SD = 10.9). Again, while a large numerical difference, it did not reach significance ($F_1 = 4.006$; *P* = .05). This corresponds to a between-group difference of 7.1 points, which exceeds the accepted threshold for a minimum clinically important difference between groups on the MADRS.[38]

In terms of percent change of MADRS score from baseline to score at study exit, the active group experienced a mean 35% change (SD = 41.1), and the active-control group experienced a mean 13.2% change (SD = 36.2); this difference was again numerically large but nonsignificant ($F_1 = 3.95$, *P* = .05).

Treatment Outcome—Rate of Change

Investigators observed a significantly faster rate of improvement in MADRS score (change in MADRS score divided by treatment weeks a patient spent in the study) in the active group compared to the active-control group. The mean change in total MADRS score per week in the active group was 1.26 points (SD = 1.63); in the active-control group, this was 0.37 points per week (SD = 0.91) ($F_1 = 4.99$; *P* = .03; $\eta^2 = 0.08$, 95% CI, 0–0.23; ANOVA).

Treatment Adherence

Patients reported high levels of treatment adherence across visits in the study on the BARS questionnaire (mean 96.4% adherence, SD = 13 in active; 95% adherence, SD = 10.9 active-control, $F_1 = 0.76$, *P* = .40).

Safety Outcomes

With respect to safety, adverse event rates (eg, medication side effects) and serious adverse event rates (eg, hospitalizations) were examined in the Safety population (48 active, 19 active-control). 89 adverse events were reported for the active group, a rate of 1.9 events per patient, compared to 51 events (2.7 events/patient) in the active-control group. There were 3 serious adverse events in the active group, and none in the active-control group. All 3 events were determined to have been unrelated to the CDSS by the site's primary investigator. Further details can be found in the Supplementary Material.

Patient Engagement

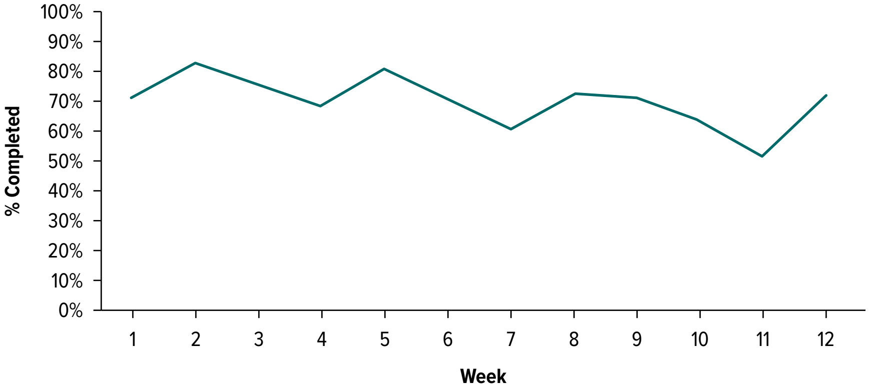
We examined PHQ-9 completion rates during the 12 treatment weeks (Figure 2). The total PHQ-9 completion rate was 70% (67% in the active group and 77% in the active-control group). These completion rates are in line with our previous feasibility study.[18,19]

Clinician Engagement

Over the course of the study, 81.25% of doctors in the active group who recruited patients accessed the CDSS at least twice (eg, at least once after training); 81.25% accessed the treatment algorithm at least twice (eg, at least once after training), and 81.25% accessed the AI

Figure 2.

PHQ-9 Completion Rates for the 12 Treatment Weeks (Total Sample n = 61)



Abbreviation: PHQ-9 = Patient Health Questionnaire-9.

Table 3.

Clinician Platform Usage by Visit

Visit	App access (%)[a]	Algorithm access (%)[b]	AI result access (%)[c]
Treatment visit 1 (Day 0)	0.88	0.69	0.67
Treatment visit 2 (W2)	0.80	0.80	0.07
Treatment visit 3 (W4–W6)	0.74	0.74	0.00
Treatment visit 4 (W8)	0.71	0.68	0.03
Treatment visit 5 (W12)	0.56	0.47	0.00

[a] App access = proportion of clinicians in the active group who logged into the app at each visit.
[b] Algorithm access = proportion of clinicians who went beyond logging in and accessed the clinical algorithm module.
[c] AI result access = proportion of clinicians who accessed the AI results at the end of 1 session of the clinical algorithm, which was intended to occur at visit 1 or 2 (1 access later in treatment was a result of an error made by a clinician and logged as a protocol deviation).
Abbreviations: AI = artificial intelligence, W = week (eg, W1 = Week 1).

results at least once (note: AI results were not available during training). Per visit, we report the number of active clinicians who accessed the application and, of these, the number that accessed the treatment algorithm and the AI results (Table 3).

DISCUSSION

This study is the first of its kind in mental health care to integrate an AI-powered CDSS in a longitudinal fashion to assist clinicians in making more effective clinical decisions about treatment selection and management while providing patients with more information about their own symptoms and trajectories. These initial results suggest that this CDSS is effective in improving remission rates (28.6% in the active group compared to no remitters in the active-control group) and the rate of symptom improvement in adult patients with moderate or greater severity depression. In addition, engagement with the platform by both clinicians and patients was high throughout the study, and no adverse events were linked to the CDSS.

Despite both groups having the same baseline depression severity, similar clinical profiles, and the same treatment options permitted, patients in the active group had significantly improved remission rates and more rapid improvement. Given this, we suggest that the CDSS may have had a positive impact on clinical decision-making, demonstrating the potential for a CDSS which can organize information, present it at the clinically appropriate time at the point of care, and provide personalized treatment outcome predictions to significantly improve outcomes.[39–41] Further discussion of potential mechanisms of action for the CDSS is presented in the Supplementary Material.

This study has several strengths. The first is a design that intended to replicate realistic use of the CDSS, where clinicians and patients were not required to use the platform or adhere to it in any particular manner. Findings of improved outcomes in the active group may therefore approximate benefits which would be derived in real clinical practice. The second key strength is the comparison of the CDSS with an active-control group that approximated realistic best practices in-clinic today, suggesting that the CDSS may be able to improve outcomes over and above these best practices. Finally, the CDSS was simple to introduce into a diverse array of clinical environments and had reasonable patient and clinician engagement, which increases its potential for rapid adoption.

This study also has several limitations. The first is the smaller than intended sample size. This limits the power of subgroup and secondary analyses, while raising the risk of statistical uncertainty, as reflected in test statistics. Small sample size studies such as this one require replication in larger samples in future work in order to ensure the consistency of the treatment effect. It is reassuring, however, that significant clinical benefit was observed in line with the a priori estimated effect size. Another significant limitation is the imbalance in the number of patients recruited into the active and control groups. On the one hand, this speaks to the interest that active group clinicians may have had in using the tool with patients (resulting in more rapid recruitment), but on the other hand a smaller control group makes comparing group outcomes more challenging. In addition, differential recruitment rates may indicate some bias—for example, that active clinicians were more motivated to recruit patients, which could have potentially translated to the quality of their treatment efforts. It is important to note, however, that patients were blinded to group allocation and were not aware of the nature of the intervention in the group they were not assigned to, which was intended to mitigate patient expectation effects. Efforts were underway to improve recruitment in the control group prior to the premature end of the study, and this would need to be prioritized in future replication studies. An additional significant limitation is the lack of remitters in the active-control group. Due to the impact of the COVID-19 pandemic, recruitment from community and primary care clinics proved challenging. As such, all patients recruited were drawn from psychiatric practices at University or Veterans Affairs affiliated sites. Therefore, the patients seen in our study were from psychiatric practices; the majority (Table 1) had recurrent depression, and there was a high comorbidity burden (Table 1). A similar sample of depressed public-sector outpatients (the majority of patients in this study were public sector) with high levels of comorbidity and recurrent depression, treated according to a medication algorithm, had remission rates of only 11% (dropping to 5.1% for sustained remission).[42] Similarly, in a pharmacogenetic decision support trial of depressed patients who had previously failed treatment,[43] the remission rate in the control group was 10%. In our study, a roughly 10% remission rate for the control group would have meant 1.9 remitters; given the small sample size, it is not surprising that by chance we did not capture remitters. Should we have had the chance to recruit more participants, we would have expected the remission rate in the control group to converge to roughly 10%. As our active group had a remission rate of nearly 30%, this would yield a difference roughly 20%, in line with our expected effect size during sample size calculation. As discussed in the methods and in investigators' companion papers,[29,30] the AI model used in this study was limited to providing initial treatment outcome predictions and could not adapt to treatment failure, and it had a preference for escitalopram being predicted as the treatment most likely to be effective, while providing more variable predictions for the other medications (see Benrimoh et al[29] for detailed discussion). In addition, first-line treatments such as psychotherapy, vortioxetine, and mirtazapine monotherapy could not be included in the AI model due to lack of data, though these were present in the treatment algorithm. Future versions of the model will continue to be improved by more and more diverse data, which will likely continue to improve the performance of the platform. Another limitation is the fact that the CDSS is a composite intervention, consisting of measurement-based care, a rule-based algorithm, and the AI model. It would be important to be able to identify, in future work, which elements of the intervention are most responsible for the clinical improvements seen.

CONCLUSIONS

In this paper, we have shown preliminary data indicating the clinical effectiveness and safety, in a cluster-randomized study, of an innovative AI-powered CDSS to support clinical decision-making in the treatment of adult patients with MDD of moderate and greater severity. Use of this and similar systems, which could be implemented rapidly into clinical practice with minimal training, has significant potential to improve the effectiveness and speed of treatment for MDD. Future work on CDSS systems like this which are intended to be integrated into longitudinal care may benefit from the study methods discussed here. Replication of this work in larger samples by independent investigators is also needed. Future work could also be directed at further expanding and improving the AI model implemented in the intervention, potentially using data collected during real-world use of the CDSS.

Article Information

Published Online: August 27, 2025. https://doi.org/10.4088/JCP.24m15634
© 2025 Physicians Postgraduate Press, Inc.

Submitted: September 26, 2024; accepted April 14, 2025.

To Cite: Benrimoh D, Whitmore K, Richard M, et al. Artificial Intelligence in Depression–Medication Enhancement (AID-ME): a cluster randomized trial of a deep-learning-enabled clinical decision support system for personalized depression treatment selection and management. *J Clin Psychiatry* 2025;86(3):24m15634.

Author Affiliations: Aifred Health Inc, Montreal, Quebec, Canada (Benrimoh, Whitmore, Richard, Golden, Perlman, Mehltretter, Fratila, Armstrong, Israel, Popescu, Anacleto, Qassim); Department of Psychiatry, McGill University, Montreal, Quebec, Canada (Benrimoh, Perlman, Looper, Beaulieu, Steiner, Margolese); Douglas Mental Health University Institute, McGill University, Verdun, Quebec, Canada (Benrimoh, Perlman, Jalali, Friesen, Barkat, Gifuni, Ferrari, Beaulieu); University of Western Ontario, London, Ontario, Canada (Golden); Department of Psychiatry, University of Arizona, Tucson, Arizona (Karp); Department of Psychiatry, University of Michigan, Ann Arbor, Michigan (Parikh); Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada (Golchi, Moodie, Shen); Department of Psychiatry, Salem Veteran Affairs Medical Center, Salem, Virginia (Sapra); Virginia Tech Carilion School of Medicine, Roanoke, Virginia (Sapra); Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada (Kloiber); Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada (Kloiber); Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada (Kloiber); Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada (Kloiber); Department of Psychiatry, Institut Universitaire en Santé Mentale de Montréal, Montreal, Quebec, Canada (Pinard); Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, Georgia (Dunlop); Yale University School of Medicine, New Haven, Connecticut (Ranganathan); Veterans Affairs Connecticut Healthcare System, West Haven, Connecticut (Ranganathan); Relief - The Path of Mental Health, Montreal, Quebec, Canada (Enault); Bipolar Disorders Clinic, Douglas Mental Health University Institute, McGill University, Verdun, Quebec, Canada (Beaulieu); Department of Psychiatry, Jewish General Hospital, Montreal, Quebec, Canada (Looper, Rej); Lady Davis Institute, Jewish General Hospital, Montreal, Quebec, Canada (Rej); School of Nursing and Midwifery, Queen's University Belfast, Belfast, Northern Ireland, United Kingdom (Rej); Department of Family Medicine, McGill University, Montreal, Quebec, Canada (Hersson-Edery); University of Waterloo, Waterloo, Ontario, Canada (Qassim); Barts and the London School of Medicine, London, England, United Kingdom (McGuire-Snieckus).

Corresponding Author: David Benrimoh, MD, CM, MSc, FRCPC, Douglas Mental Health University Institute, 6875 Blvd. LaSalle, Verdun, QC H4H 1R3, Canada (david.benrimoh@mcgill.ca).

Author Contributions: Conceived the study and revised the paper: (Benrimoh, Israel, Perlman, Richard, Margolese); conducted and supervised the study and analyses and wrote the first draft of the paper: (Benrimoh); assisted with data collection and manuscript review: (Whitmore, Richard, Golden, Israel, Perlman, Popescu, Anacleto, Qassim); assisted with analysis and paper revision: (Whitmore, Richard, Golden, Perlman, Jalali, Friesen, Barkat, Fratila, Popescu); produced the machine learning model used in the study and revised the paper: (Mehltretter, Fratila, Armstrong, Perlman, Benrimoh); served on the steering committee and revised the paper: (Karp, Parikh, Golchi, Moodie, Enault, Rej, Hersson-Edery, Margolese); served as statisticians who oversaw sample size calculations carried out by Shen and reviewed the statistical analysis plan, which was approved by Benrimoh and Margolese: (Golchi, Moodie); served as site primary investigators and revised the paper: (Gifuni, Ferrari, Parikh, Sapra, Kloiber, Pinard, Dunlop, Looper, Ranganathan); served on the data safety monitoring board and revised the paper: (Beaulieu, Steiner); provided the STAR questionnaire and revised the paper: (Snieckus); provided supervision for the study, the analysis, and revised the paper (Margolese).

Relevant Financial Relationships: Drs Benrimoh and **Perlman**; **Mss Whitmore**, **Richard**, **Golden**, **Armstrong**, **Israel**, **Popescu**, **Anacleto**, and **Qassim**; and **Messrs Mehltretter** and **Fratila** were employees and/or shareholders of Aifred Health Inc. and supported this research in the context of their work for Aifred Health. **Drs Karp** and **Parikh** are members of Aifred Health's scientific advisory board. **Dr Parikh** has received honoraria from Aifred Health. **Drs Benrimoh**, **Golchi**, **Moodie**, and **Rej** receive a salary award from the Fonds de recherche du Québec – Santé (FRQS). **Dr Moodie** is a Canada Research Chair (Tier 1) in Statistical Methods for Precision Medicine. **Dr Dunlop** has received research support from Boehringer Ingelheim, Compass Pathways, Intra-Cellular Therapies, NIMH, Otsuka, and Usona Institute and has served as a consultant for Biohaven, Cerebral Therapeutics, Myriad Neuroscience, and Otsuka. **Dr Rej** receives grant funding from Mitacs (for a graduate student), is on a steering committee for Abbvie, and owns shares of Aifred Health. **Dr Kloiber** reports grants from the Labatt Family Innovation Fund in Brain Health (Department of Psychiatry, University of Toronto), the Max Bell Foundation, the Canadian Centre on Substance Use and Addiction, the Centre for Addiction and Mental Health Discovery Fund, the Ontario Ministry of Health and Long-Term Care (MOHLTC), the Canadian Institutes of Health Research (CIHR), and the International OCF Foundation (IOCDF) and received honorarium for consultation from EmpowerPharm. **Dr Karp** has been provided options in Aifred Health. **Dr Margolese** has received honoraria, sponsorship, or grants for his participation as a consultant, advisory committee member, and/or as a speaker at educational events for AbbVie, HLS Therapeutics Inc, Janssen, Lundbeck, Otsuka, Newron, Sunovion, and Teva and has received grants and/or research support from the MGH Hospital Foundation, SyneuRX, and Aifred Health. **Ms Jalali** and **Messrs Friesen** and **Barkat** received honoraria from Aifred Health for working on this paper. The other authors report no relevant conflicts.

Funding/Support: Funding was provided by Aifred Health, MEDTEQ FSISSS, Bell Let's Talk and Brain Canada, Investissement Quebec, the Quebec Ministry of Economy and Innovation, the Mindstrong Foundation at the Jewish General Hospital, and the McGill Industry and Partnership grant.

Role of the Sponsors: The study was codesigned and supervised by Dr Margolese and Aifred Health. No other funder had a role in the development or reporting of this research.

Acknowledgments: We thank patients, physicians, and local staff for participating in this study.

Additional Information: The full protocol and statistical analysis plan as well as the supplementary material are available from the authors on request from Dr Benrimoh at david.benrimoh@mail.mcgill.ca.

Code Availability: The methods used to develop the AI platform are available in previous publications; however, the final code for the CDSS and final trained model are not publicly available at this time. Previous versions of the pipeline used to construct the model are available here (https://github.com/Aifred-Health/VulcanAI), and a full version of the model trained using a pharmaceutical dataset is available here (https://github.com/Aifred-Health/pharma_research_model).

Data Availability: Data requests will be considered by the study sponsor for noncommercial applications, on the condition that this publication is referenced in publications using the dataset. Requests for data access should be directed to Dr Benrimoh at david.benrimoh@mcgill.ca and will be responded to within 30 days. Data requested will be transferred in a deidentified format using patient ID numbers.

Supplementary Material: Available on request from the corresponding author.

References

1. Greenberg PE, Fournier A-A, Sisitsky T, et al. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). *J Clin Psychiatry*. 2015;76(2):155–162.
2. World Health Organization. *Depression and Other Common Mental Disorders: Global Health Estimates*. World Health Organization; 2017.
3. Rush AJ, Trivedi MH, Wisniewski SR, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. *Am J Psychiatry*. 2006;163(11):1905–1917.
4. Kraus C, Kadriu B, Lanzenberger R, et al. Prognosis and improved outcomes in major depression: a review. *Transl Psychiatry*. 2019;9(1):127.
5. Benrimoh D, Fratila R, Israel S, et al. Aifred Health, a deep learning powered clinical decision support system for mental health. In: *The NIPS'17 Competition: Building Intelligent Systems*. Springer International Publishing; 2018:251–287.
6. Mehltretter J, Rollins C, Benrimoh D, et al. Analysis of features selected by a deep learning model for differential treatment selection in depression. *Front Artif Intell*. 2019;2:31.
7. Squarcina L, Villa FM, Nobile M, et al. Deep learning for the prediction of treatment response in depression. *J Affect Disord*. 2021;281:618–622.
8. Poon AIF, Sung JJY. Opening the black box of AI-Medicine. *J Gastroenterol Hepatol*. 2021;36(3):581–584.
9. Maslej MM, Kloiber S, Ghassemi M, et al. Out with AI, in with the psychiatrist: a preference for human-derived clinical decision support in depression care. *Transl Psychiatry*. 2023;13(1):210.
10. Celi LA, Cellini J, Charpignon ML, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities-A global review. *PLOS Digit Health*. 2022;1(3):e0000022.
11. Schneider F, Kratz S, Bermejo I, et al. Insufficient depression treatment in outpatient settings. *Ger Med Sci*. 2004;2:Doc01.
12. Lisinski A, Hieronymus F, Eriksson E, et al. Low SSRI dosing in clinical practice-a register-based longitudinal study. *Acta Psychiatr Scand*. 2021;143(5):434–443.
13. von Knorring J, Baryshnikov I, Jylhä P, et al. Prospective study of antidepressant treatment of psychiatric patients with depressive disorders: treatment adequacy and outcomes. *BMC Psychiatry*. 2023;23(1):888.
14. Golden G, Popescu C, Israel S, et al. Applying artificial intelligence to clinical decision support in mental health: what have we learned? *Health Pol Technology*. 2022;13(2):100844.
15. Kennedy SH, Lam RW, McIntyre RS, et al. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of

Adults with Major Depressive Disorder: Section 3. Pharmacological Treatments. *Can J Psychiatry*. 2016;61(9):540–560.
16. Benrimoh D, Tanguay-Sela M, Perlman K, et al. Using a simulation centre to evaluate preliminary acceptability and impact of an artificial intelligence-powered clinical decision support system for depression treatment on the physician-patient interaction. *BJPsych Open*. 2021;7(1):e22.
17. Tanguay-Sela M, Benrimoh D, Popescu C, et al. Evaluating the perceived utility of an artificial intelligence-powered clinical decision support system for depression treatment using a simulation center. *Psychiatry Res*. 2022;308:114336.
18. Popescu C, Golden G, Benrimoh D, et al. Evaluating the clinical feasibility of an artificial intelligence-powered, web-based clinical decision support system for the treatment of depression in adults: longitudinal feasibility study. *JMIR Form Res*. 2021;5(10):e31862.
19. Qassim S, Golden G, Slowey D, et al. A mixed-methods feasibility study of a novel AI-enabled, web-based, clinical decision support system for the treatment of major depression in adults. *J Affect Disord Rep*. 2023;14:100677.
20. Mehltretter J, Fratila R, Benrimoh D, et al. Differential treatment Benet prediction for treatment selection in depression: a deep learning analysis of STAR*D and CO-MED data. *Comput Psychiatr*. 2020;4:61.
21. Kleinerman A, Rosenfeld A, Benrimoh D, et al. Treatment selection using prototyping in latent-space with application to depression treatment. *PLoS One*. 2021;16(11):e0258400.
22. Benrimoh D, Kleinerman A, Furukawa TA, et al. Towards outcome-driven patient subgroups: a machine learning analysis across six depression treatment studies. *Am J Geriatr Psychiatry*. 2024;32(3):280–292.
23. Liu X, Rivera SC, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ*. 2020;370:m3164.
24. Cook AJ, Delong E, Murray DM, et al. Statistical lessons learned for designing cluster randomized pragmatic clinical trials from the NIH Health Care Systems Collaboratory Biostatistics and Design Core. *Clin Trials*. 2016;13(5):504–512.
25. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)*. American Psychiatric Publishing; 2013.
26. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry*. 1998;59(suppl 20):22–57.
27. Asberg M, Montgomery SA, Perris C, et al. A comprehensive psychopathological rating scale. *Acta Psychiatr Scand Suppl*. 1978(271):5–27.
28. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. 2001;16(9):606–613.
29. Benrimoh D, Armstrong C, Mehltretter J, et al. Development and validation of a deep-learning model for differential treatment benefit prediction for adults with major depressive disorder deployed in the Artificial Intelligence in Depression Medication Enhancement (AIDME) study. arXiv [q-bio.NC]; 2024. https://doi.org/10.48550/arXiv.2406.04993
30. Perlman K, Mehltretter J, Benrimoh D, et al. Development of a differential treatment selection model for depression on consolidated and transformed clinical trial datasets. *Transl Psychiatry*. 2024;14(1):263.
31. Fleischhacker WW, Czobor P, Hummer M, et al. Placebo or active control trials of antipsychotic drugs? *Arch Gen Psychiatry*. 2003;60(5):458–464.
32. Center for Drug Evaluation and Research. *E10 choice of control group and related issues in clinical trials*. U.S. Food and Drug Administration; 2020. Accessed February 8, 2024. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e10-choice-control-group-and-related-issues-clinical-trials.
33. Byerly MJ, Nakonezny PA, Rush AJ. The Brief Adherence Rating Scale (BARS) validated against electronic monitoring in assessing the antipsychotic medication adherence of outpatients with schizophrenia and schizoaffective disorder. *Schizophr Res*. 2008;100(1–3):60–69.
34. McIntyre RS, Konarski JZ, Mancini DA, et al. Measuring the severity of depression and remission in primary care: validation of the HAMD-7 scale. *CMAJ*. 2005; 173(11):1327–1334.
35. Kaneriya SH, Robbins-Welty GA, Smagula SF, et al. Predictors and moderators of remission with aripiprazole augmentation in treatment-resistant late-life depression: an analysis of the IRL-GRey randomized clinical trial. *JAMA Psychiatry*. 2016;73(4):329–336.
36. Page SJ, Persch AC. Recruitment, retention, and blinding in clinical trials. *Am J Occup Ther*. 2013;67(2):154–161.
37. Williams JBW, Kobak KA. Development and reliability of a structured interview guide for the Montgomery Asberg Depression Rating Scale (SIGMA). *Br J Psychiatry*. 2008;192(1):52–58.
38. Hengartner MP, Plöderl M. Estimates of the minimal important difference to evaluate the clinical significance of antidepressants in the acute treatment of moderate-to-severe depression. *BMJ Evid Based Med*. 2022;27(2):69–73.
39. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115–118.
40. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402–2410.
41. He J, Baxter SL, Xu J, et al. The practical implementation of artificial intelligence technologies in medicine. *Nat Med*. 2019;25(1):30–36.
42. Rush AJ, Trivedi M, Carmody TJ, et al. One-year clinical outcomes of depressed public sector outpatients: a benchmark for subsequent studies. *Biol Psychiatry*. 2004;56(1):46–53.
43. Greden JF, Parikh SV, Rothschild AJ, et al. Impact of pharmacogenomics on clinical outcomes in major depressive disorder in the GUIDED trial: a large, patient- and rater-blinded, randomized, controlled study. *J Psychiatr Res*. 2019;111: 59–67.