Title: 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

Authors: David Benrimoh*[1,2,3,8], Kate Whitmore[1], Maud Richard[1], Grace Golden[1,4], Kelly Perlman[1,3], Sara Jalali[8], Timothy Friesen[8], Youcef Barkat[8], Joseph Mehltretter[1], Robert Fratila[1], Caitrin Armstrong[1], Sonia Israel[1], Christina Popescu[1], Jordan F. Karp[5], Sagar V. Parikh[6], Shirin Golchi[7], Erica EM Moodie[7], Junwei Shen[7], Anthony J. Gifuni[8], Manuela Ferrari[8], Mamta Sapra[9,10], Stefan Kloiber[11,12,13,14], Georges-F. Pinard[15], Boadie W. Dunlop[16], Karl Looper[2,21], Mohini Ranganathan[17,18], Martin Enault[19], Serge Beaulieu[2,20], Soham Rej[21,22,23], Fanny Hersson-Edery[24], Warren Steiner[2], Alexandra Anacleto[1], Sabrina Qassim[1,4], Rebecca McGuire-Snieckus[25], Howard C. Margolese[2]

Affiliations:

1. Aifred Health Inc., Canada, Quebec, Montreal, H3J 1M1
2. Department of Psychiatry, McGill University, Canada, Quebec, Montreal, H3A 0G4
3. McGill University, Canada, Quebec, Montreal, H3A 0G4
4. University of Western Ontario, Canada, ON, London, N6A 3K7
5. Department of Psychiatry, University of Arizona, United States, Arizona, Tucson, 85721
6. Department of Psychiatry, University of Michigan, United States, Michigan, Ann Arbour, 48109
7. Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Canada, Quebec, Montreal, H3A 0G4
8. Douglas Mental Health University Institute, McGill University, Canada, Quebec, Verdun, H4H 1R3
9. Department of Psychiatry, Salem Veteran Affairs Medical Center, United States, Virginia, Salem, 24153
10. Virginia Tech Carilion School of Medicine, United States, Virginia, Roanoke, 24016
11. Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Canada, Ontario, Toronto M5T 1R8
12. Department of Psychiatry, University of Toronto, Canada, Ontario, Toronto, M5S 1A1
13. Institute of Medical Science, University of Toronto, Canada, Ontario, Toronto, M5S 1A1
14. Department of Pharmacology and Toxicology, University of Toronto, Canada, Ontario, Toronto, M5S 1A1
15. Department of Psychiatry, Institut Universitaire en Santé Mentale de Montréal, Canada, Quebec, Montreal, H1N 3M5
16. Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, United States, Georgia, Atlanta, 30322
17. Yale University School of Medicine, United States, Connecticut, New Haven, 06510
18. Veterans Affairs Connecticut Healthcare System, United States, Connecticut, West Haven, 06516
19. Relief - The Path of Mental Health, Canada, Quebec, Montreal, H2L 1J6
20. Bipolar Disorders Clinic, Douglas Mental Health University Institute, McGill University, Canada, Quebec, Verdun, H4H 1R3
21. Department of Psychiatry, Jewish General Hospital, Canada, Quebec, Montreal, H3T 1E2
22. Lady Davis Institute, Jewish General Hospital, Canada, Quebec, Montreal, H3T 1E2
23. School of Nursing and Midwifery, Queen's University Belfast, United Kingdom, Northern Ireland, Belfast, BT9 7BL
24. Department of Family Medicine, McGill University, Canada, Quebec, Montreal H3A 0G4
25. Barts and the London School of Medicine, United Kingdom, England, London, E1 2AD

*Corresponding Author:

David Benrimoh, MD.CM., MSc., MSc., FRCPC
Douglas Mental Health University Institute
Phone: +1 (514) 463-7813
Email: david.benrimoh@mail.mcgill.com

Abstract:

Major Depressive Disorder (MDD) is a leading cause of disability and there is a paucity of tools to personalize and manage treatments. A cluster-randomized, patient-and-rater-blinded, clinician-partially-blinded study was conducted to assess the effectiveness and safety of the Aifred Clinical Decision Support System (CDSS) facilitating algorithm-guided care and predicting medication remission probabilities using clinical data. Clinicians were randomized to the Active (CDSS access) or Active-Control group (questionnaires and guidelines access). Primary outcome was remission (<11 points on the Montgomery Asberg Depression Rating Scale (MADRS) at study exit). Of 74 eligible patients, 61 (42 Active, 19 Active-Control) completed at least two MADRS (analysis set). Remission was higher in the Active group (n = 12/42 (28.6%)) compared to Active-Control (0/19 (0%)) (p = 0.01, Fisher's exact test). No adverse events were linked to the CDSS. This is the first effective and safe longitudinal use of an artificial intelligence-powered CDSS to improve MDD outcomes.

Keywords:
Major depression, clinical decision support system, artificial intelligence, machine learning

Introduction

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 valuable to have a scalable point-of-care tool which can help personalize treatment choice without requiring expensive testing[56].

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[7]. Most studies have differentiated between two treatments (e.g. two drugs) or treatment types, (two 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"[8,9-10]. In addition, 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 which supports clinicians in the implementation of guidelines (2016 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.

Results

Sites

Ten sites were recruited and were cleared to recruit patients; of these 1 site was closed early because of lack of capacity to complete the trial. 8 sites recruited patients into the study. Sites were located in Canada (5) and the United States (4) and included U.S.Veterans Affairs hospitals and mood disorders programs in university-affiliated psychiatric departments.

Recruitment - Clinicians

50 clinicians were recruited, consistent with the recruitment target. 26 were randomized in the Active group and 24 in the Active-Control group. 39 of these clinicians were psychiatrists; 2 were nurse practitioners specialized in psychiatry, and 9 were psychiatry residents. 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 one patient (57%); 16 in the Active group (64%) and 11 in the Active-Control group (50%). Active and Active-Control clinicians spent essentially the same mean number of months in the study (Active = 9.9 months; Active-Control = 10.1 months). Further details are available in the Supplementary Material.

Recruitment and Dropout - Patients

Patients were recruited between 2022-06-15 and 2023-11-16, 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 = 0.15, Fisher's exact test)). Considering only patients who attended visit 1 (V1), 36/48 (75%) of Active patients completed all 12 weeks, and 18/19 (95%) completed visit 5 (V5); this was not statistically significant (p = 0.09, Fisher's exact test). Further details can be found in the Supplementary Material.

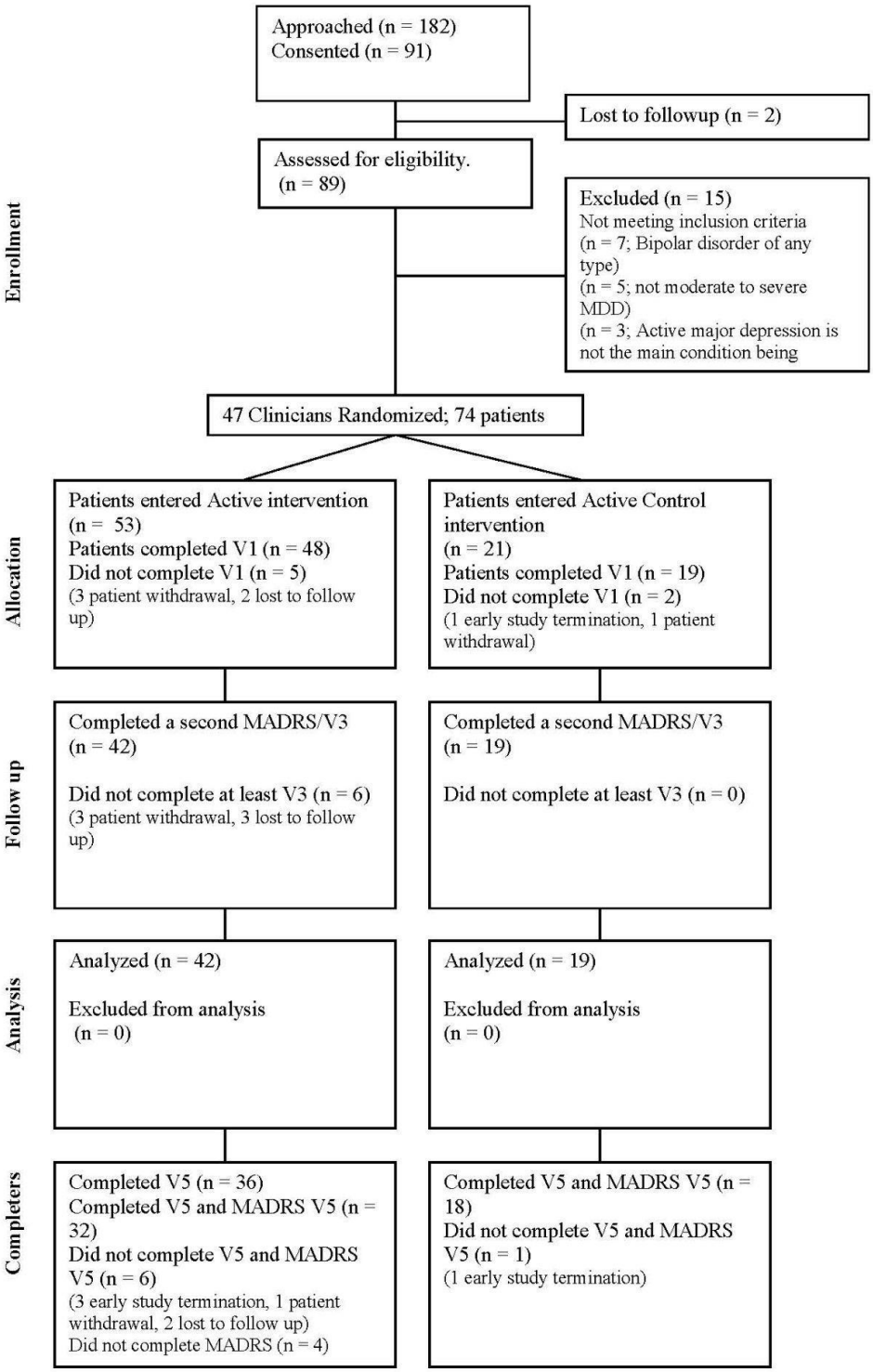

Figure 1. Patient Consort Diagram

Demographics and baseline clinical characteristics

Intervention and Active-Control groups did not differ with regards 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.

	Active group (n = 42)	Active-Control group (n = 19)
Mean age (SD)	44.0 (15.2)	39.3 (12.4)
Sex - n female (%)	20 (47.6)	10 (52.6)
Mean baseline MADRS (SD)	33 (7.3)	30 (5.8)
Race	n = 40	n = 18
- White	30 (75.0)	13 (72.2)
- Other**	10 (25.0)	5 (27.8)
Mean yearly household income (USD)	n = 37	n = 15
- Mean (SD)	42,206.87 (29103.3)	44,817.10 (17975.4)
Highest level of education achieved*	n = 40	n = 17
- Some high school/ high school diploma or equivalent (GED)	5 (15)	5 (27.8)
- Some university or college	12 (30.0)	4 (22.2)
- Bachelor's degree	11 (27.5)	6 (33.3)
- Graduate or Professional Degree	7 (17.5)	1 (5.6)
- Trade/technical training or other	5 (12.5)	1 (5.6)
Graduated high school	39 (97.5)	17 (94.4)
Currently employed	n = 39	n = 18
	18 (41.5)	12 (66.7)
Marital Status		
- Single	22 (55.0)	12 (66.7)

- Partnered	18 (45.0)	6 (33.3)
Mean number of medications (all indications) prescribed at baseline (n = 57)	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 months	6 (14.3)	6 (31.6)
- Agoraphobia current	5 (12)	2 (10.5)
- Substance Use Disorder (Non-Alcohol), Past 12 months	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)

Table 1: Baseline Clinical and Demographic Characteristics per Group
*Note: participants could select more than one option; the graduated high school entry was constructed based on the available data.
** Lower count rows have been collapsed into the 'other' category in order to preserve confidentiality
Further demographic details are available in the Supplementary Materials.

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 = 0.01, 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 ($X^2 = 3.6$, p = 0.06). The proportion of responders was not significantly different at visit 3 or 5, but was significantly different at visit 4 (p = 0.04, 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 = 0.05). 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 [23].

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 non-significant (F(1) = 3.95, p = 0.05).

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 Active-Control. 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 = 0.03; eta-squared = 0.08, 95% CI [0,0.23], ANOVA). See Figure 2.

Outcome	Active	Active-Control	p-value
Remission	12 (28.6%)	0 (0%)	0.01
Response	17 (40.5%)	3 (15.8%)	0.06
Mean change from baseline	12 (SD = 13.5)	4.9 (SD = 10.9)	0.05
Percent change from baseline	35 (SD = 41.1)	13.2 (36.2)	0.05
Slope of improvement (amount of change per week)	1.26 (SD = 1.63)	0.37 (SD = 0.91)	0.03

Table 2: Summary of Outcomes

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 = 0.40).

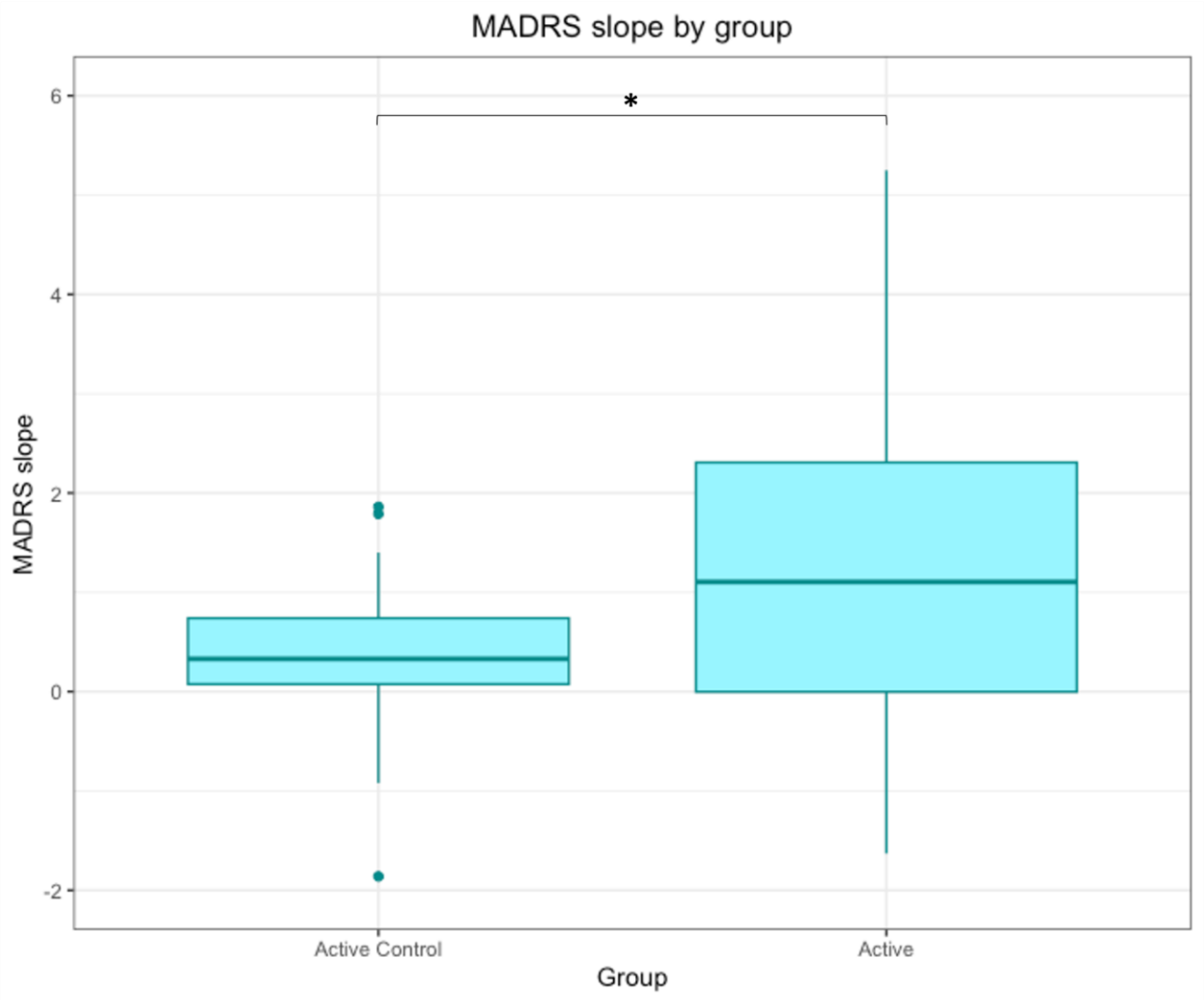

Figure 2: MADRS Slope by Group - bar plot displaying significantly different mean MADRS slope per group (F(1) = 4.99; p = 0.03; eta-squared = 0.08, 95% CI [0,0.23], ANOVA) Error bars represent standard deviation.

Safety Outcomes

With respect to safety, adverse event rates (e.g., medication side effects) and serious adverse event rates (e.g., hospitalizations) were examined in the Safety population (all patients who completed at least the first treatment visit; this included 48 Active and 19 Active-Control patients). With respect to adverse events, 89 were reported for the Active group, a rate of 1.9 events per patient. 51 adverse events were reported in the Active-Control group, a rate of 2.7 events per patient. As such, the intervention was not associated with an increase in adverse event rate. There were 3 serious adverse events in the Active group, and none in the Active-Control group. All three events were determined to have been unrelated to the CDSS by the site's primary investigator. The three events included a visit to a psychiatric emergency department because of suicidal ideation; a brief hospitalization in a psychiatric short stay unit because of a panic attack and suicidal ideation; and a visit to general emergency department for a suicidal attempt gesture (overdose with 6 pills of 60mg duloxetine). All patients recovered from these events and none required prolonged hospitalization. As all three patients were known for cluster B personality traits or borderline personality disorder, the site PIs categorized these as expected events. Further details of the events and clinician perception of application safety can be found in the Supplementary Material.

Patient Engagement

Patient engagement was primarily determined by their responses to questionnaires sent to them regularly in the application, on a weekly or bi-weekly basis. Taking the PHQ-9, which was sent weekly, as an exemplar, investigators examined the completion rates during the 12 treatment weeks (Fig. 3). The total PHQ-9 completion rate was 70% (67% in active, 77% in the Active-Control group). These completion rates are in line with investigators' previous feasibility study[18,19]. Completion rates for the other regular application questionnaires can be found in the Supplementary Material.

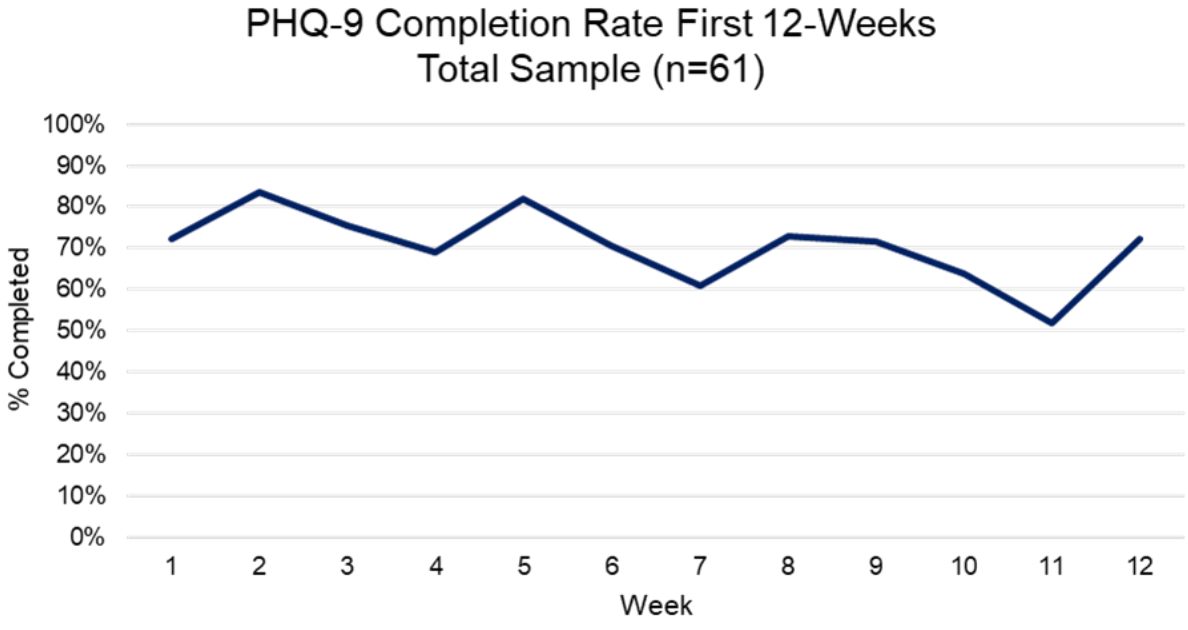

Figure 3: PHQ-9 completion rates for the 12 treatment weeks.

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 (e.g. at least once after training); 81.25% accessed the treatment algorithm at least twice (e.g. at least once after training), and 81.25% accessed the AI results at least once (note: AI results were not available during training). Per visit, investigators reported the number of Active clinicians who accessed the application; of these, the number that accessed the treatment algorithm and the AI results were reported. Of note, the majority of AI result accesses happened at visit 1 or 2, as expected, as the AI results were intended to be generated early in treatment (one access later in treatment was a result of an error made by a clinician and logged as a protocol deviation).

Visit	App Access (%)	Algorithm Access (%)	AI Result Access (%)
Treatment Visit 1 (Day 0)	0.88	0.69	0.67
Treatment Visit 2 (W2)	0.80	0.80	0.07
Treatment Visit 3 (W4-6)	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

Table 3: Clinician Platform Usage by Visit App access = proportion of clinicians in the active group who logged into the app at each visit; Algorithm Access = proportion of clinicians who went beyond logging in and access the clinical algorithm module; AI Result Access = proportion of clinicians who accessed the AI results at the end of one session of the clinical algorithm, which was intended to occur at Visit 1 or 2.

Discussion

This study is the first of its kind in mental healthcare 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. Investigators have demonstrated that this CDSS is safe and 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 majority of the study, and no adverse events were linked to the CDSS. Questionnaire completion rates were consistent with what investigators observed in feasibility testing[18], and the rate at which Active clinicians chose AI-consistent treatments was close to the consistent treatment rate estimated in a previous simulation center study [16].

In this study, no experimental treatments were introduced, the same treatments were permitted in both groups, and patients were treated by their own clinician who was free to use or discard the information provided by the CDSS as they saw fit. In addition, patients had access to the same platform in both treatment groups. Despite both groups having the same baseline depression severity, and the same treatment options permitted, patients in the Active group had significantly improved remission rates and more rapid improvement. Given this, investigators suggest that it is likely that the CDSS had a positive impact on clinical decision-making, and potentially shared decision-making between clinicians and patients. The study therefore demonstrates 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 the treatment of patients who experience significant suffering and whose illness can generate significant societal and healthcare costs. Results echo those of AI decision support studies in other areas of medicine, where AI powered tools have been shown to be potential augmentors of clinical decision making [24–26]. Further discussion of potential mechanisms of action for the CDSS is presented in the Supplementary Material.

It is important to discuss the potential generalizability of these results. Sites were diverse in nature and demographics show that patients were diverse in terms of their backgrounds and comorbidities. All treating clinicians were psychiatrists, psychiatry residents, or specialized nurse practitioners. This occurred despite best efforts to recruit primary care providers, which proved to be difficult due several factors (lack of embedded research staff, time-consuming clinic-level onboarding in busy clinical practices, and primary care service adaptation to the post COVID-19 environment). However, previous work has shown that the CDSS is feasible in primary care [19], [18], and AI training data included patients in both primary and specialized services [22]. Given that the majority of MDD is treated by primary care physicians, and that patients in primary care are more likely to have less treatment resistant or recurrent depression, future work will need to confirm similar if not improved results in primary care[27]. Indeed, previous work demonstrated that primary care clinicians found the CDSS to be more useful than psychiatrists did; given the complexity of MDD management, the CDSS could be a valuable tool in primary care [16,17]. In addition, given the lack of safety concerns identified in this and previous studies [18,19], it would seem reasonable to introduce the CDSS to primary care in future work. Finally, while the clinical algorithm based on the guidelines might be applicable across many jurisdictions, the AI model was trained on data mostly from European and North American populations and as such would need to be validated and potentially re-trained before being used outside of these populations given potentially different patterns of symptom expression in different cultures [28–30].

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. Robust findings of improved outcomes in the Active group are therefore likely indicative of benefits which would be derived in real clinical practice. The second major strength is the comparison of the CDSS with an Active-Control group that approximated realistic best practices in-clinic today, which suggests that the CDSS will 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 high 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 was caused by early study termination due to resource restrictions resulting from delays related to COVID-19. This limits the power of subgroup and secondary analyses. It is reassuring, however, that significant clinical benefit was observed in line with the *a priori* estimated effect size. As discussed in the methods and in investigators' companion paper (Perlman et al. 2024), 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[31] for detailed discussion). 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 which elements of the intervention are most responsible for the clinical improvements seen. This could have been accomplished by

adding further arms to the trial, but this was not possible. To compensate, the design matched the Active and Active-Control interventions closely, such that the main differences between groups were the clinical algorithm and the AI predictions. The clinical algorithm, in turn, was approximated in the Active-Control group by the guideline training and questionnaire data provided to clinicians. The objective of the study was to determine the impact of the CDSS as a unitary intervention; future implementation research could focus on separating the platform's component parts in order to study their independent effects. Another limitation is the imbalance in the number of patients recruited into the Active and Active-Control groups. This may speak to the interest that Active group clinicians may have had in using the tool with patients (resulting in more rapid recruitment). Efforts were underway to improve recruitment in the Active-Control group prior to the premature end of the study.

Conclusions

In this paper, investigators demonstrate 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. 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. Further analyses based on this dataset will examine qualitative and quantitative data about clinician and patient perceptions of the platform. Similar methods as the ones presented here could potentially be used to assess clinical decision support in other disease areas.

Registration at clinicaltrials.gov: NCT04655924

The full protocol and statistical analysis plan are available in the Supplementary Material.

Funding: 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. The study was co-designed and supervised by HM and Aifred Health. No other funder had a role in the development or reporting of this research.

DISCLOSURES

Declaration of Competing Interest:

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

DB, KW, MR, GG, KP, JM, RF, CA, SI, CP, AA, and SQ were employees and/or shareholders of Aifred Health Inc. and supported this research in the context of their work for Aifred Health. JFK, SP are members of Aifred Health's scientific advisory board. SP has received honoraria from Aifred Health. DB, SG, EEMM, and SR receive a salary award from the Fonds de recherche du Québec – Santé (FRQS). EEMM is a Canada Research Chair (Tier 1) in Statistical Methods for Precision Medicine. BWD has received research support from Boehringer Ingelheim, Compass Pathways, Intra-Cellular Therapies, NIMH, Otsuka, Usona Institute, and has served as a consultant for Biohaven, Cerebral Therapeutics, Myriad Neuroscience, and Otsuka. SR receives

grant funding from Mitacs (for a graduate student), is on a steering committee for Abbvie, and owns shares of Aifred Health. SK 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). SK received honorarium for consultation from EmpowerPharm. JFK has been provided options in Aifred Health. HCM 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. He has received grants and/or research support from the MGH Hospital Foundation, SyneuRX and Aifred Health. SJ, TF, and YB recieved honoraria from Aifred Health for working on this paper. All other authors report no relevant conflicts.

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). Code describing model evaluation of generation of results is available in the Supplementary Material.

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

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

Authorship statement:
DB, SI, KP, MR, HM, conceived the study and revised the paper. DB conducted and supervised the study and analyses, and wrote the first draft of the paper. KW, MR, GG, SI, KP, CP, AA, and SQ assistant with data collection and manuscript review. KW, MR, GG, KP, SJ, TF, YB, RF and CP assisted with analysis and paper revision. JM, RF, CA, KP, and DB produced the machine learning model used in the study and revised the paper. JK, SP, SG, EM, ME, SR, FHE, and HM served on the steering committee and revised the paper. SG and EM served as statisticians who oversaw sample size calculations carried out by JS and reviewed the statistical analysis plan, which was approved by DB, MR and HM. AG, MF, SP, MS, SK, GP, BD, KL, MR, served as site primary investigators and revised the paper. SB, WS served on the data safety monitoring board and revised the paper. RMS provided the STAR questionnaire and revised the paper. HM provided supervision for the study, the analysis, and revised the paper.

Online Methods:

This study is reported as per the CONSORT-AI checklist [32]. The study 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. The study was conducted in accordance with Good Clinical Practice. Written informed consent was obtained from all study participants.

Design

The current study is a two-arm, cluster-randomized trial, with clinicians serving as the cluster. Clinicians were allowed to recruit a maximum of 10 patients in order to reduce the impact of within physician intra-class correlation on inference. The expected cluster size was 7. Clinicians rather than patients were randomized as they were the ones receiving the decision support intervention, and to avoid contamination [33]. Patients entered the intervention arm of their treating clinician. No changes were made to study design after study initiation other than 1) to allow clinicians to participate longer than the originally planned 9 months in order to reach recruitment targets and 2) once the need for early study termination was determined, all participants currently enrolled were invited to complete exit interviews rather than being randomly selected.

Participants - Clinicians

Clinicians were recruited by the site primary investigator and treated as research participants for the purpose of the study, signed consent forms, and were overseen by site primary investigators. Clinicians could include primary care doctors, psychiatrists, residents overseen within their primary residency program by a participating clinician, nurse practitioners (with or without specialized mental health training), or nurse practitioner students overseen by a participating nurse practitioner. They needed to see at least one 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. This was done as the CDSS is intended to be helpful at scale. Patients were recruited from the practices or hospital-based clinics of the participating clinicians. Inclusion criteria were as follows: 1) age 18 and over 2) diagnosed by their treating clinician with MDD using DSM-5 criteria [34] 3) MDD diagnosis confirmed via a blinded rater who completed the Mini Neuropsychiatric Interview (MINI) [35] and 4) at least moderate severity, as assessed by a blinded rater completing the Montgomery Asberg Depression Rating Scale (using a cutoff of 20) [36]. Patients must have been able to 5) provide their own informed consent and 6) needed to agree to be treated by their clinician for depression, understanding that they might use a range of approved treatments which might be presented in the CDSS, and understanding they were able to provide or withhold consent for any particular treatment. Contraception was used as per usual clinical practice. Exclusion criteria were as follows: 1) age under age 18; 2) presence of bipolar disorder of any type (e.g. by clinician diagnosis or identified on the MINI) or in the patient history (except in cases where the history of bipolar disorder in the medical record was vague and was not confirmed by repeat clinical assessment or the MINI); 3) inability or unwillingness to give informed consent; 4) inability to manage patient safely as an outpatient (importantly, patients with suicidal ideation who were deemed safe to be managed as outpatients by the treating clinicians were eligible); 5) an active major depressive disorder was not the main condition being treated; and 6) an inability to use the tool (e.g. 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. In addition, psychiatric comorbidities (aside from bipolar disorder) were permitted. There were no inclusion or exclusion criteria related to input data to the AI.

Settings

Eligible settings included any public or private outpatient setting in the United States or Canada which provided outpatient care for patients presenting with MDD, with the exception that highly specialized settings where treatments were generally pre-determined (such as ketamine clinics) were not approached for participation. Both primary care and psychiatric services were invited to participate. Given these broad requirements, a diverse array of sites joined the study. These included public sector psychiatric clinics in Canada and university-affiliated and Veteran's Affairs mental health services in the U.S. Participating sites included: the Douglas Mental Health University Institute, the McGill University Health Centre, the Jewish General Hospital, the Centre for Addiction and Mental Health, Michigan University, Emory University, the Salem VA Health Care System and VA Connecticut Health Care.

Intervention - Aifred CDSS

The Aifred CDSS platform consists of the following elements. 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. The *clinician portal*, accessible by web browser, allows clinicians to see all the information patients enter, while at the same time giving them access to the *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 (PHQ-9) [37], and current treatments. For example, if a patient has been on a treatment that has not resulted in early improvement (25% decrease in baseline PHQ-9 score) after four weeks, the algorithm will remind clinicians that the guidelines recommend treatment switch (or augmentation if, for example, this is not the first treatment trial). The algorithm provides new information at each patient visit, based on patient progress. The clinician portal and its clinical algorithm are mainly focused on solving the treatment management problem and do not utilize AI. The AI component 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 two 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. 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 was available. The AI is a deep learning model trained on 9042 patients from depression treatment trials, with remission as the training objective. It takes symptom and demographic questionnaire-based data as inputs, with both the clinician and the patient each completing a short, dedicated questionnaire at the beginning of treatment to provide this data (the AI-related or "custom" questionnaires). When responding to these questionnaires, full responses are required, meaning that there was no missing data when AI predictions were made; all questions were previously validated standardized questions, a decision investigators made in order to improve data reliability. The input data is based on feature selection performed during model training [38] aidme model paper]. Patients responding to the patient version of the custom questionnaire did not require any expertise, as they were responding to questions previously validated for patient self-report.

Clinicians responding to the clinician version of the custom questionnaire (which consisted of 6 questions from the HAM-D scale and one from the HAM-A scale [39] were provided with interview guides and written prompts to help them ask and respond to these questions, but in keeping with the naturalistic design of the trial were not provided separate training on how to administer these questionnaires. The model predictions are intended to be used to inform the first treatment initiation or change in the study; the model does not update predictions after treatment failure. In order to improve interpretability, a list of the 5 most important item responses for each treatment's prediction are displayed. In order to help the clinician understand the remission probability in context, they are provided with the baseline probability of remission in the training data; the patient's mean probability of remission across predicted treatments (i.e. the average of the 10 predicted probabilities), which provides a sense of initial overall treatment success likelihood for the patient; and the relative increase or decrease relative to the patient's mean remission probability for each treatment (which helps clinicians get a sense of the ranking of each treatment relative to the others). The choice to present remission *probability*, rather than a class prediction (e.g. remit or non-remit) was made in order to provide clinicians with more nuanced information and avoid an overly-prescriptive approach to AI that would infringe on clinician autonomy[40]. The version of the AI model used in this study is extensively described in these publications [38], aidme model paper].

A visual depiction of the Aifred CDSS is available in the Supplementary Material.

Intervention - Patients

All patients received access to the same patient portal of the CDSS, where they were able to respond to questionnaires, track their responses over time using graphs, and enter and track their current and past treatments. Patients were trained to use the platform by study staff Patients did not have access to the AI or the clinical algorithm. Patient experiences differed only in their interaction with their clinicians, who had different information available based on their group assignment. Patients remained in the study for 12 weeks from their first treatment visit, and were required to see their clinician, in person or via telemedicine, at week 2, weeks 4-6, week 8 and week 12.

Intervention - Clinicians

There were two intervention groups: an Active group and an Active-Control group (hereinafter referred to as the Active-Control group)[41,42]. The Active-Control group was provided with all the tools required to perform best-practice measurement-based and guideline-informed care[15]. Clinicians in the Active-Control were provided with the results of questionnaires patients completed as well as training on the guidelines[15]. Guideline training involved a powerpoint presentation by DB 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. They were provided with training on the CDSS and on how to interpret AI results (i.e. probabilities of remission). They were instructed to consider their clinical judgment and the limitations of the AI (for example, that it only considered the data provided in the AI-related questionnaires; that it did not adapt after treatment failure; that the training data had limited demographic features available, meaning that a thorough assessment of social determinants of health was warranted for every patient; and that the model, consistent

with the training data, would usually rank escitalopram as the most effective treatment, while providing more variable rankings for the other treatments (see[31]) when making treatment decisions. Active clinicians 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.

As all raw data provided to clinicians was the same in the Active and Active-Control groups, 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.

Measures

At baseline, patients were asked to complete a demographics questionnaire as well as several clinical questionnaires. These included the Mini International Neuropsychiatric Interview (MINI) and Montgomery–Åsberg Depression Rating Scale (MADRS), assessed by the blinded rater; Patient Health Questionnaire (PHQ-9), self-report Quick Inventory of Depressive Symptomatology (QIDS-SR-16, depression), General Anxiety Disorder (GAD-7, anxiety), Alcohol Use Disorders Identification Test (AUDIT, alcohol use disorder), Drug Abuse Screen Test (DAST-10, drug use), Self-Administered Standardized Assessment of Personality – Abbreviated Scale (SAPAS-SA, personality disorder screening), Adverse Childhood Experiences (ACE), Life Events Checklist for DSM-5 (LEC-5) (both assess trauma history), and the World Health Organization Disability Assessment Schedule (WHODAS V 2.0, disability assessment) all via self-report. They were asked to complete the patient AI-related custom questionnaire no more than 2 weeks prior to their first visit, in order to ensure the AI results were reflective of their current condition. They were also asked to complete a PHQ-9 and GAD-7 weekly once they had accounts on the CDSS.

The MADRS was administered by the blinded rater at screening (no more than 2 weeks prior to the first treatment visit), 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 [43].

Clinicians were asked to complete the clinician version of the AI "custom questionnaire" at visit 1 (on paper in the Active-Control, and in the CDSS in the Active group). All clinicians were asked to complete a post appointment questionnaire within 48 hours of each patient visit detailing patient safety information, their assessment of patient status, any treatment changes made, and, for Active clinicians, their impression of the CDSS.

Outcomes

The pre-specified primary outcome of the study was remission of depressive symptoms, defined as a score of <11 [44,45] on the MADRS at study exit for those patients with at least two MADRS scores.Remission was chosen as the main outcome as it is the outcome which guidelines recommend [15]. 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. Subgroup analyses aimed at examining whether patients

receiving AI-prediction consistent treatments had improved outcomes and other secondary and exploratory analyses are available in the Supplementary.

Sample Size

We proceeded with the effect size calculation for a cluster randomized trial[46]. The intracluster correlation coefficient was set at 0.05[47]. The baseline remission value was set at 35%, based on studies in similar populations[3,48]. Cluster size was set at 7, and minimum effect size to detect was set at a 20% difference in remission rate. This was intended to be a conservative estimate based on machine learning results (see[6,20–22] and previous studies using measurement based care and algorithm-guided treatment which found larger differences in remission rate[49,50]. At 90% power, these parameters generated a requirement for 47 clinicians and 325 patients. Investigators aimed to recruit 350 patients and up to 50 clinicians. No interim analyses were planned.

Randomization, Sequence Generation, Allocation Concealment, and Implementation

Randomization of clinicians proceeded at 1:1 to the Active and Active-Control groups. Randomization was stratified by clinician type: primary care clinicians and non-specialized nurse practitioners were coded as being “primary care” and psychiatrists and nurse practitioners with mental health specialization were coded as being “specialized care”. Block randomization with a block size of 4 was used. Sequence generation proceeded using cluster randomization where clinicians were randomized and any patients allocated to those clinicians were assigned to the same arm as the clinician. This randomization was programmed and performed in SAS. Allocation was concealed until the participating clinician completed enrollment procedures by using an interactive web response system. The sequence was generated by Hong Chen at Alimentiv Inc. and then retrieved by local site coordinators after clinician enrollment who then informed clinicians of their intervention group.

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 and that there were two groups. It was explained that while their experience of the platform in each group would be the same, and that they would provide the same information, the clinicians in each group would use the information in different ways. They were told that their clinician would explain how they were using the information. Clinicians were instructed not to reveal group allocations or to tell patients what the clinicians in the other group were provided with.

Clinicians

Clinicians were aware of their group assignments as they were the ones receiving the AI predictions and investigators judged that, at this early stage in clinical AI research, providing clinicians with fake predictions would have been ethically questionable. Clinicians were instead partially blinded in the following manner to reduce expectation bias[51]: they were not told the study endpoints, and they were not informed of the expected effect sizes of the interventions.

Raters

Raters who collected the primary outcome (MADRS) and conducted the MINI were blind to group allocation.

Study Staff

Study staff were not blind to group allocation as they needed to provide technical support to the clinicians in the study, conduct study interviews, and provide clinicians in the Active-Control with the weekly questionnaires. Researchers conducting the analysis and those supporting the sites were likewise not blinded to group allocation.

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 two ratings of the MADRS (the Analysis set). Safety data were analyzed for the Safety population, pre-specified as all patients who attended at least the first treatment visit. Missing data were not imputed. Analyses were carried out using SPSS (IBM) version 29.0.1.1, Microsoft Excel, and R studio. Data were collected using the Aifred CDSS (Aifred Health) and TrialStat (TrialStat). Those conducting analysis were not blinded.

Demographic and baseline clinical data were summarized and are presented in Table 1. Baseline MADRS was compared between groups using one-way ANOVA. The primary outcome (MADRS remission) was assessed using a Fisher's exact test due to the lack of remitters in the Active-Control group. As pre-specified, sensitivity analyses were carried out (see Supplementary Material). Secondary outcomes were compared using one-way ANOVAs, and proportions were compared using two-sided X^2 or two-sided Fisher's exact tests, as appropriate. Cox models were initially to be used to assess time to remission; as this was not possible because of the lack of remitters in the control group, this analysis was replaced with an analysis of slope of MADRS change (change in score over time in study). While the analysis plan called for adjustment by clinician type, this was not relevant or necessary as only specialist clinicians were recruited. To assess patient engagement, investigators calculated the percentage of the self-report questionnaires completed by patients during the 12 treatment weeks, accounting for study dropout. Physician engagement was determined by examining clinician access logs for the platform at each visit. Secondary and exploratory analyses were not corrected for multple comparisons. Further pre-specified Supplementary analyses are detailed in the Supplementary Material.

Early Study Termination

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

REFERENCES:

1. Greenberg, P. E., Fournier, A.-A., Sisitsky, T., Pike, C. T. & Kessler, R. C. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). *J. Clin. Psychiatry* **76**, 155–162 (2015).
2. Health Organization, W. Depression and other common mental disorders: global health estimates. (2017).
3. Rush, A. J. *et al.* Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. *Am. J. Psychiatry* **163**, 1905–1917 (2006).
4. Kraus, C., Kadriu, B., Lanzenberger, R., Zarate, C. A., Jr & Kasper, S. Prognosis and improved outcomes in major depression: a review. *Transl. Psychiatry* **9**, 127 (2019).
5. Benrimoh, D. *et al.* Aifred Health, a Deep Learning Powered Clinical Decision Support System for Mental Health. in *The NIPS '17 Competition: Building Intelligent Systems* 251–287 (Springer International Publishing, 2018).
6. Mehltretter, J. *et al.* Analysis of Features Selected by a Deep Learning Model for Differential Treatment Selection in Depression. *Front Artif Intell* **2**, 31 (2019).
7. Squarcina, L., Villa, F. M., Nobile, M., Grisan, E. & Brambilla, P. Deep learning for the prediction of treatment response in depression. *J. Affect. Disord.* **281**, 618–622 (2021).
8. Poon, A. I. F. & Sung, J. J. Y. Opening the black box of AI-Medicine. *J. Gastroenterol. Hepatol.* **36**, 581–584 (2021).
9. Maslej, M. M., Kloiber, S., Ghassemi, M., Yu, J. & Hill, S. L. Out with AI, in with the psychiatrist: a preference for human-derived clinical decision support in depression care. *Transl. Psychiatry* **13**, 210 (2023).
10. Celi, L. A. *et al.* Sources of bias in artificial intelligence that perpetuate healthcare disparities-A global review. *PLOS Digit Health* **1**, e0000022 (2022).
11. Schneider, F. *et al.* Insufficient depression treatment in outpatient settings. *Ger. Med. Sci.* **2**, Doc01 (2004).

12. Lisinski, A., Hieronymus, F., Eriksson, E. & Wallerstedt, S. M. Low SSRI dosing in clinical practice-a register-based longitudinal study. *Acta Psychiatr. Scand.* **143**, 434–443 (2021).
13. von Knorring, J. *et al.* Prospective study of antidepressant treatment of psychiatric patients with depressive disorders: treatment adequacy and outcomes. *BMC Psychiatry* **23**, 888 (2023).
14. Golden, G. *et al.* Applying artificial intelligence to clinical decision support in mental health: What have we learned? *Health Policy and Technology* 100844 (2024).
15. Kennedy, S. H. *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* **61**, 540–560 (2016).
16. Benrimoh, D. *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* **7**, e22 (2021).
17. Tanguay-Sela, M. *et al.* Evaluating the perceived utility of an artificial intelligence-powered clinical decision support system for depression treatment using a simulation center. *Psychiatry Res.* **308**, 114336 (2022).
18. Popescu, C. *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 formative research* **5**, e31862 (2021).
19. Qassim, S. *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. *Journal of Affective Disorders Reports* **14**, 100677 (2023).
20. Mehltretter, J. *et al.* Differential treatment Benet prediction for treatment selection in depression: A deep learning analysis of STAR*D and CO-MED data. *Comput. Psychiatr.* **4**, 61 (2020).
21. Kleinerman, A. *et al.* Treatment selection using prototyping in latent-space with application

to depression treatment. *PLoS One* **16**, e0258400 (2021).

22. Benrimoh, D. *et al.* Towards Outcome-Driven Patient Subgroups: A Machine Learning Analysis Across Six Depression Treatment Studies. *Am. J. Geriatr. Psychiatry* **32**, 280–292 (2024).
23. Hengartner, M. P. & 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* **27**, 69–73 (2022).
24. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
25. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402–2410 (2016).
26. He, J. *et al.* The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **25**, 30–36 (2019).
27. McIntyre, R. S., Millson, B. & Power, G. S. Burden of Treatment Resistant Depression (TRD) in patients with major depressive disorder in Ontario using Institute for Clinical Evaluative Sciences (ICES) databases: Economic burden and healthcare resource utilization. *J. Affect. Disord.* **277**, 30–38 (2020).
28. Kessler, R. C. & Bromet, E. J. The epidemiology of depression across cultures. *Annu. Rev. Public Health* **34**, 119–138 (2013).
29. Kirmayer, L. J., Jarvis, G. E. & Gómez-Carrillo, A. Depression across cultures: An ecosocial approach. *Textbook of mood disorders* (2021).
30. Krendl, A. C. & Pescosolido, B. A. Countries and Cultural Differences in the Stigma of Mental Illness: The East–West Divide. *J. Cross. Cult. Psychol.* **51**, 149–167 (2020).
31. Benrimoh, D. *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).

32. Liu, X. *et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ* **370**, m3164 (2020).
33. Cook, A. J., Delong, E., Murray, D. M., Vollmer, W. M. & Heagerty, P. J. Statistical lessons learned for designing cluster randomized pragmatic clinical trials from the NIH Health Care Systems Collaboratory Biostatistics and Design Core. *Clin. Trials* **13**, 504–512 (2016).
34. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. (American Psychiatric Publishing, 2013).
35. Sheehan, D. V. *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* **59 Suppl 20**, 22–33;quiz 34–57 (1998).
36. Asberg, M., Montgomery, S. A., Perris, C., Schalling, D. & Sedvall, G. A comprehensive psychopathological rating scale. *Acta Psychiatr. Scand. Suppl.* 5–27 (1978).
37. Kroenke, K., Spitzer, R. L. & Williams, J. B. The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* **16**, 606–613 (2001).
38. Perlman, K. *et al.* Development of a differential treatment selection model for depression on consolidated and transformed clinical trial datasets. *medRxiv* (2024) doi:10.1101/2024.02.19.24303015.
39. Maier, W., Buller, R., Philipp, M. & Heuser, I. The Hamilton Anxiety Scale: reliability, validity and sensitivity to change in anxiety and depressive disorders. *J. Affect. Disord.* **14**, 61–68 (1988).
40. Benrimoh, D. *et al.* Editorial: ML and AI safety, effectiveness and explainability in healthcare. *Front. Big Data* **4**, 727856 (2021).
41. Fleischhacker, W. W. *et al.* Placebo or active control trials of antipsychotic drugs? *Arch. Gen. Psychiatry* **60**, 458–464 (2003).
42. Center for Drug Evaluation & Research. E10 choice of control group and related issues in

clinical trials. *U.S. Food and Drug Administration* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e10-choice-control-group-and-related-issues-clinical-trials (2020).

43. Byerly, M. J., Nakonezny, P. A. & Rush, A. J. 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.* **100**, 60–69 (2008).
44. McIntyre, R. S. *et al.* Measuring the severity of depression and remission in primary care: validation of the HAMD-7 scale. *CMAJ* **173**, 1327–1334 (2005).
45. Kaneriya, S. H. *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* **73**, 329–336 (2016).
46. Rutterford, C., Copas, A. & Eldridge, S. Methods for sample size determination in cluster randomized trials. *Int. J. Epidemiol.* **44**, 1051–1067 (2015).
47. Underwood, M. *et al.* Exercise for depression in elderly residents of care homes: a cluster-randomised controlled trial. *Lancet* **382**, 41–49 (2013).
48. Rush, A. J. *et al.* Combining medications to enhance depression outcomes (CO-MED): acute and long-term outcomes of a single-blind randomized study. *Am. J. Psychiatry* **168**, 689–701 (2011).
49. Guo, T. *et al.* Measurement-Based Care Versus Standard Care for Major Depression: A Randomized Controlled Trial With Blind Raters. *Am. J. Psychiatry* **172**, 1004–1013 (2015).
50. Adli, M. *et al.* How Effective Is Algorithm-Guided Treatment for Depressed Inpatients? Results from the Randomized Controlled Multicenter German Algorithm Project 3 Trial. *Int. J. Neuropsychopharmacol.* **20**, 721–730 (2017).
51. Page, S. J. & Persch, A. C. Recruitment, retention, and blinding in clinical trials. *Am. J. Occup. Ther.* **67**, 154–161 (2013).

52. McGuire-Snieckus, R., McCabe, R., Catty, J., Hansson, L. & Priebe, S. A new scale to assess the therapeutic relationship in community mental health care: STAR. *Psychol. Med.* **37**, 85–95 (2007).

53. Popescu, C. *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.* **5**, (2021).

54. Perlman, K. *et al.* A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. *J. Affect. Disord.* **243**, 503–515 (2019).