Psychiatry Research 308 (2022) 114336

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Evaluating the perceived utility of an artificial intelligence-powered clinical decision support system for depression treatment using a simulation center

Myriam Tanguay-Sela [a,#], David Benrimoh [a,#,*], Christina Popescu [a], Tamara Perez [b], Colleen Rollins [c], Emily Snook [d], Eryn Lundrigan [b], Caitrin Armstrong [a], Kelly Perlman [b], Robert Fratila [a], Joseph Mehltretter [a], Sonia Israel [a], Monique Champagne [a], Jérôme Williams [b], Jade Simard [e], Sagar V. Parikh [f], Jordan F. Karp [g], Katherine Heller [h], Outi Linnaranta [i], Liliana Gomez Cardona [i], Gustavo Turecki [i], Howard C. Margolese [b]

[a] *Aifred Health Inc., Montreal, Quebec, Canada*
[b] *McGill University, Montreal, Quebec, Canada*
[c] *University of Cambridge, Cambridge, England, UK*
[d] *University of Toronto, Toronto, Ontario, Canada*
[e] *Université du Québec à Montréal, Montreal, Quebec, Canada*
[f] *University of Michigan, Ann Arbor, Michigan, United States*
[g] *University of Arizona, Tucson, Arizona, United States*
[h] *Duke University, Durham, North Carolina, United States*
[i] *Douglas Mental Health University Institute, McGill University, Verdun, Quebec, Canada*

ARTICLE INFO

Key Words:
Major depressive disorder
Artificial intelligence
Primary care
Outpatient treatment
Physician-patient relationship
Simulation center
Patient-Physician Relationship

ABSTRACT

Aifred is a clinical decision support system (CDSS) that uses artificial intelligence to assist physicians in selecting treatments for major depressive disorder (MDD) by providing probabilities of remission for different treatment options based on patient characteristics. We evaluated the utility of the CDSS as perceived by physicians participating in simulated clinical interactions. Twenty physicians who were either staff or residents in psychiatry or family medicine completed a study in which they had three 10-minute clinical interactions with standardized patients portraying mild, moderate, and severe episodes of MDD. During these scenarios, physicians were given access to the CDSS, which they could use in their treatment decisions. The perceived utility of the CDSS was assessed through self-report questionnaires, scenario observations, and interviews. 60% of physicians perceived the CDSS to be a useful tool in their treatment-selection process, with family physicians perceiving the greatest utility. Moreover, 50% of physicians would use the tool for all patients with depression, with an additional 35% noting that they would reserve the tool for more severe or treatment-resistant patients. Furthermore, clinicians found the tool to be useful in discussing treatment options with patients. The efficacy of this CDSS and its potential to improve treatment outcomes must be further evaluated in clinical trials.

1. Introduction

Clinical decision support systems (CDSS) consolidate large quantities of clinical information to provide clinicians with actionable data to support medical decision-making and assist with managing treatment protocols (Sutton et al., 2020; Zikos and DeLellis, 2018). Increasingly, artificial intelligence (AI) algorithms are being integrated into CDSS, allowing for the deployment of predictive analytics by clinicians as part of routine practice (Sutton et al., 2020).

Depression is the leading cause of disability worldwide (World Health Organization, 2017), and more than one in nine people will experience depression over the course of their lives (Bromet et al., 2011). Despite its high prevalence and the availability of many effective treatment modalities, only a third of patients will reach remission after their first treatment (Warden et al., 2007) and there are few widely available tools to help physicians select the optimal treatment for each

[*] Corresponding author.
E-mail address: david.benrimoh@mail.mcgill.ca (D. Benrimoh).
[#] Myriam Tanguay-Sela and David Benrimoh contributed equally to this work.

https://doi.org/10.1016/j.psychres.2021.114336
Received 21 April 2021; Received in revised form 6 December 2021; Accepted 9 December 2021
Available online 11 December 2021

patient. Thus, antidepressants are frequently prescribed using trial and error, an often lengthy process which can cause a great deal of patient suffering (Benrimoh et al., 2018). An artificial intelligence-powered CDSS enables physicians to identify treatments with an increased likelihood of remission for each individual patient, thus minimizing the number of unsuccessful treatment trials.

We created a CDSS which integrates an AI model that uses individual patient clinical and demographic characteristics to provide clinicians with remission probabilities for specific depression treatments accompanied by clinical practice guidelines (Benrimoh et al., 2021). Previous computerized decision support tools for depression have included screening assistance and treatment guidelines. For example, they can help physicians determine when to adjust medication doses, start augmentation treatments, or change treatment course (Harrison et al., 2020; Trivedi et al., 2004; Rollman et al., 2002). While these tools are helpful to manage depression, they provide similar information when compared to existing best practice guidelines, and do not offer personalized treatment suggestions based on individual patient characteristics without relying on expensive or time-consuming tests (e.g. imaging or genetic testing). Our CDSS offers a novel approach, using individual patient clinical and demographic variables to provide personalized remission predictions which are integrated directly into best practice guidelines.

We studied our CDSS at a simulation center to evaluate its ease of use in routine practice, impact on the physician-patient interaction, and perceived utility. The two former points have been addressed in a previous publication (Benrimoh et al., 2021) and the latter is the focus of this paper. We were interested in how useful primary care physicians and psychiatrists would find the CDSS in assisting with their treatment decisions and in discussing these treatment options with patients. To our knowledge, the perceived utility of this kind of tool has not been addressed in previous studies of depression treatment.

The CDSS design was informed by important characteristics identified in discussions with physicians, including the simplicity of the interface and clinical utility of information displayed, as well as integration of the AI results into existing guidelines. It was important that the integration of the AI into the CDSS followed generalizable principles that, if validated, could be used to design similar tools. This included layering the AI predictions on top of existing best evidence guidelines (the CANMAT guidelines of the treatment of depression: Kennedy et al., 2016) and clearly labeling the AI predictions so that clinicians could consciously choose when to use them in their decision-making. To improve interpretability (Benrimoh et al., 2018) and preserve physician autonomy, we designed our AI tool to provide reports detailing the key variables that informed each prediction and displayed the AI results as probabilities of remission, rather than suggestions or positive recommendations (see Mehltretter et al., 2020, for methodological details). The tool was also designed to facilitate shared decision-making as a means of fostering patient autonomy and agency in the clinical encounter while positively supporting the clinician-patient relationship. Full details of the methodology followed in developing the CDSS, how it is differentiated from other CDSS, as well as the design of the simulation center study, can be found in Benrimoh et al. (2021).

Our main hypothesis was that clinicians would perceive the tool to be useful in shared decision-making with patients. We also hypothesized that primary care physicians (PCPs) and psychiatrists would use the tool differently due to their differing expertise and experience in the treatment of depression.

2. Methods

2.1. Participants and study design

The study sample consisted of intended users of the CDSS: staff and residents specialized in primary care or psychiatry. Twenty participating physicians were recruited for the study via social media and email. Participants provided informed consent and were compensated for their time. The study was approved by the Douglas Mental Health University Institute Research Ethics Board.

The study was conducted at the McGill Steinberg Centre for Simulation and Interactive Learning. The simulation center provided nine standardized patients (SPs), professional actors trained to act as patients, who were compensated for their involvement. It should be noted that the use of SPs has been shown to be a valid reflection of patient experience (Beullens et al., 1997; Shirazi et al., 2011). SPs also ensure internal validity by allowing for each participant to respond to identical clinical scenarios. The simulation center setup included a one-way mirror arrangement and an auditory monitoring system, allowing research assistants (RAs) to observe and listen to the simulated clinical scenarios.

Three 10-minute clinical scenarios of mild, moderate and severe depression were created by a clinician (D.B.), based on real patient data from the de-identified datasets on which the CDSS model was trained. Clinican participants experienced the scenarios in a random order. Further details on the scenarios can be found in the supplementary methods and in Benrimoh et al. (2021).

Before the start of the simulation, we gave a short presentation to introduce participants to the basic principles of AI that our model uses. This was a structured presentation which can be found in the supplementary methods and included an explanation of the type of data used to train the model, the fact that a neural network was used, as well as the model metrics and the output of the model. Participants did not have significant prior knowledge of the AI or its development process. Participants then underwent a 10-minute training session with an RA which included teaching on navigating the tool. They were informed that the 'patients' had used the tool to complete questionnaires prior to their session, but that they had a limited understanding of how the AI model operated. Clinicians had access to standardized questionnaire results (the PHQ-9 (Kroenke et al., 2001), QIDS-SR-16 (Rush et al., 2003), and the HAM-D (Hamilton, 1960) in the application. The questions used by the AI to generate predictions, which included questions from the HAM-D, QIDS-SR-16 and IDSC (Rush et al., 2000), as well as some other symptom and demographic questions selected by the AI, are presented in the supplementary materials.

Participants conducted the 10-minute clinical consultations with the SP as per their usual practice and integrated the CDSS as they saw fit. It was suggested, but not mandated, that they spend five minutes interviewing the patient followed by five minutes using the CDSS. Within the CDSS, participants had access to the patient's questionnaire results and the treatment algorithm with the integrated predictive model. The provided laptop was positioned at a 45° angle towards the participant in order to make the screen visible to researchers observing the interaction, though participants were free to move the laptop as preferred. Participants were informed that the clinical scenarios were 10 minutes long, and had access to a clock. Following each clinical scenario, participants completed a questionnaire regarding their use of the CDSS model. After the three scenarios, physician participants were interviewed by RAs using a standard semi-structured interview with predefined questions including both open-ended and specific questions (see supplementary methods) and completed a questionnaire about their experience using the model, as well as a short quiz to assess their familiarity with the CANMAT 2016 guidelines for depression treatment. SPs were interviewed in an unstructured manner as a group at the end of each of the three testing days and their observations were recorded in writing by RAs. This was done in order to obtain their immediate impressions about their interactions with clinicians. In addition, given that SPs sometimes changed between testing days, this ensured that all SPs were able to provide feedback.

To improve reporting quality, we endeavored to report our results in line with the suggested amendments to the STROBE guidelines for the reporting of simulation-based research (Cheng et al., 2016). Custom questionnaires were used due to the novelty of the CDSS. Additionally, participants were surveyed about their previous simulation experience.

2.2. Quantitative analysis

Our quantitative analysis was primarily aimed at providing descriptive results derived from three questionnaires completed by participants. The first was a demographics questionnaire, the results of which are available in Table 1. The second was a custom-created questionnaire administered to participants at the end of their simulation experience; this was intended to capture a number of different aspects of their overall experience using the tool with the SPs and is the source of the data for Figs. 1 and 2. The third questionnaire was another custom questionnaire administered immediately after each of the three simulated patient sessions, intended to capture their immediate feelings after each session, and is the source of the data for Figs. 3 and 4. The questionnaires had to be designed in a custom manner in order to capture the novel aspects of using this technology. In this paper, we present results for the questions related to perceived utility; these questions can be found in the supplementary materials. The descriptive results from participant self-report questionnaires were generated using R v.3.3.2 and visualized using the ggplot2 v.3.2.1 package.

2.3. Qualitative analysis

The qualitative data consisted of the written and interview feedback from the participants, SPs and the RAs' written observations on the clinical scenarios. Interviews were not recorded, instead RAs took extensive notes, which they transcribed into digital spreadsheets, where data was then coded. Initial themes were brainstormed prior to data analysis and interpretation to mitigate bias while coding the data. Nine themes emerged from this initial effort: 1) interpretability of the AI report; 2) degree of trust in AI; 3) user experience for the patient; 4) user experience for the clinician; 5) the tool's impact on the physician-patient interaction; 6) the potential role and use of the AI tool in practice; 7) suggested tool improvements; 8) user interface; and 9) situations in which the clinician felt they would use the tool. Each theme had a number of related sub-themes. Using an inductive thematic analysis approach allowed the data to be coded without trying to fit it into a preexisting coding frame or the researcher's analytic preconceptions, and allowed the themes identified to be strongly linked to the data themselves (Braun and Clarke, 2006).

Table 1
Demographics (n= 20).

Demographic	Response Options	Frequency
Gender	**Female**	**13 (65%)**
	Male	**7 (35%)**
Specialty	**Primary Care Physician (PCP)**	**9 (45%)**
	Psychiatrist	**11 (55%)**
Training	**Quebec**	**17 (85%)**
Where were you trained?	**Rest of Canada**	**3 (15%)**
Resident Level	**N/A (staff clinicians) = 14**	
If you are a resident, what is your level?	**PGY1**	**1 (5%)**
	PGY2	**2 (10%)**
	PGY3	**3 (15%)**
Staff Years Experience	**Residents**	**6 (30%)**
If you are a staff member, how many years of experience do you have?	**0–5**	**4 (20%)**
	6–10	**2 (10%)**
	11–15	**2 (10%)**
	16–20	**4 (20%)**
	21+	**2 (10%)**
Environments of Practice	**Hospital Outpatient**	**10 (50%)**
Which environment(s) do you practice in? (Check all that apply.)	**Department**	**8 (40%)**
	Inpatient Service	**12 (60%)**
	Outpatient Clinic	**4 (20%)**
	ER	**1 (5%)**
	Consult Liaison	**1 (5%)**
	Specialized Mood Disorder Service	
Patients with MDD Treated per Month	**<10**	**6 (30%)**
How often would you say you treat people with major depressive disorder (MDD) (number of patients per month)?	**11–20**	**7 (35%)**
	21–30	**2 (10%)**
	31–40	**2 (10%)**
	41–50	**1 (5%)**
	51–60	**2 (10%)**

The investigator triangulation method was employed during the qualitative analysis, in which multiple investigators compared individually coded qualitative data to reduce bias (Archibald, 2016). RAs independently read and coded excerpts of the data into subheadings of the thematic table. The source of the excerpt and the scenario from which it was extracted was noted. Four RAs were each assigned the data corresponding to 10 participants, such that each participant's data was independently coded by two RAs. An additional independent coder compiled all the excerpts. This data was condensed and redundancies eliminated by collapsing some of the themes and rearranging the subheadings. The RAs then independently reread and coded all of the data into a final summary table. This stage ensured that any data that had been missed in earlier coding stages could be added, and also validated the new themes in relation to the full data set. Triangulation of qualitative data sources involved the comparison of observational data with interview data, allowing the analysis of different perspectives. This approach offered greater insight into the relationship between inquiry approach, data sources, and the phenomena under study.

3. Results

3.1. Quantitative analysis

3.1.1. Sample description

Our sample comprised 11 psychiatrists (8 staff, 3 residents) and 9 PCPs (6 staff, 3 residents). The mean age of participants was 39.5 years (SD 13.3). Most of the participants were trained in Quebec (n= 17), and ranged from residents to staff with decades of experience (see Table 1). Physicians' clinical practice environments included both inpatient and outpatient services, and there was a wide range in the self-estimated number of patients treated for MDD per month (Table 1). The responses to questions related to current clinical practices, the potential role for clinical decision aids and AI technologies, as well as which treatments were prescribed during the simulation sessions can be found in Supplementary Table 1.

3.2. Descriptive results

Question responses were recorded using Likert scales with the answer options "strongly disagree", "somewhat disagree", "unsure", "somewhat agree", and "strongly agree". Here, "disagreed to a degree" indicates that the "somewhat" and "strongly" disagree options have been combined; and "agreed to a degree" or "to some degree" indicates that the "somewhat" and "strongly" results have been combined.

At study end, 60% of participants agreed to some degree that the model was useful in making treatment decisions (for all participants, mean score 3.6 where "strongly disagree" is coded as "1″ and "strongly agree" is coded as "5″, SD 1.1; for psychiatrists, mean 3.2, SD 1.0; for PCPs, mean 4.0, SD 1.2). More PCPs appeared to feel that the model was useful in making treatment decisions compared to psychiatrists (Fig. 1).

Tests of statistical significance are not included for these figures because of the small sample size which means the study is underpowered compared to the number of comparisons shown. These figures are meant to help assess possible trends which should be replicated in larger samples in future studies.

At study end, 70% of participants described the probabilities produced by the model as "reasonable", while 15% described them as "too pessimistic" and the remaining 15% as "too optimistic" (Fig. 2). There did not appear to be substantial divergence between PCP and

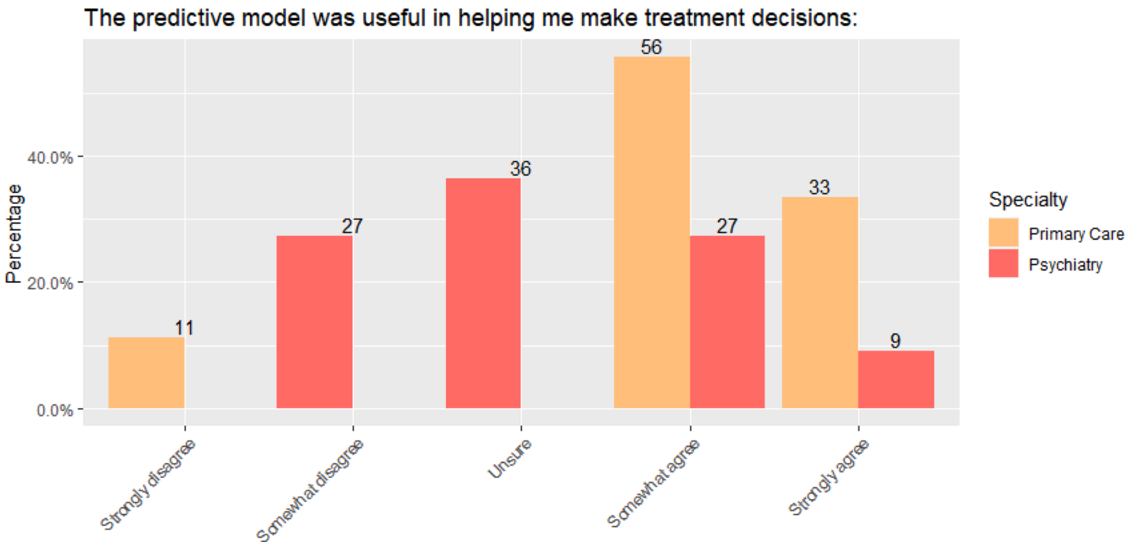

Fig. 1. Ratings of the usefulness of the model in making treatment decisions.

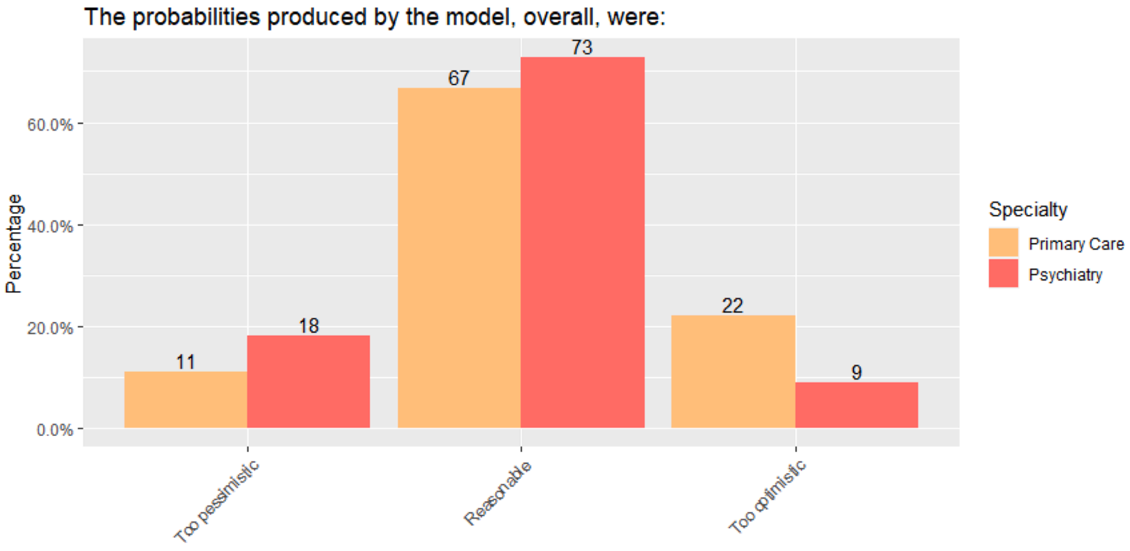

Fig. 2. Ratings of perceived reasonableness of the model.

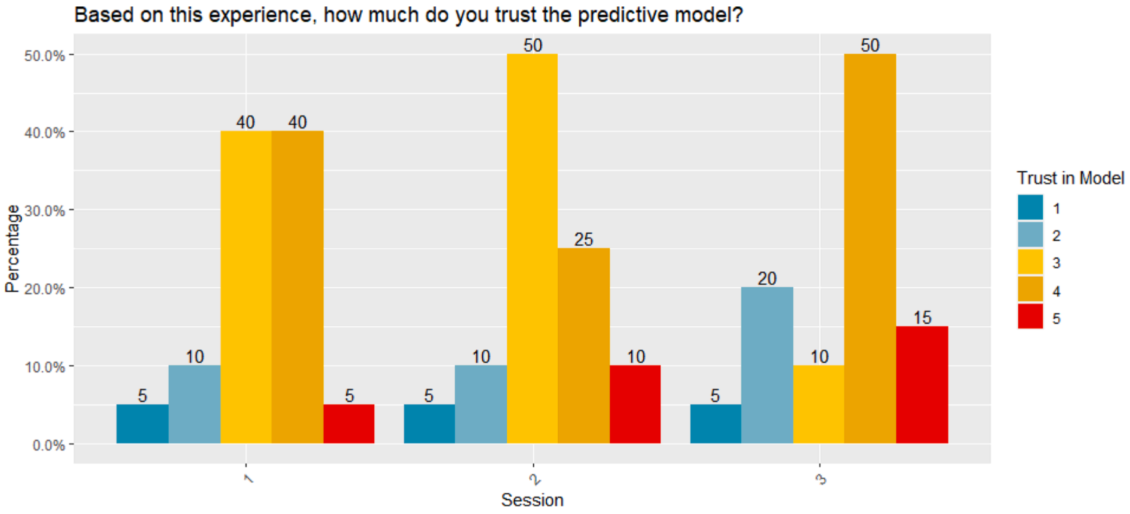

Fig. 3. Ratings of trust in the model across simulation sessions (scale of 1 to 5, 1 being "very little" and 5 being "very much").

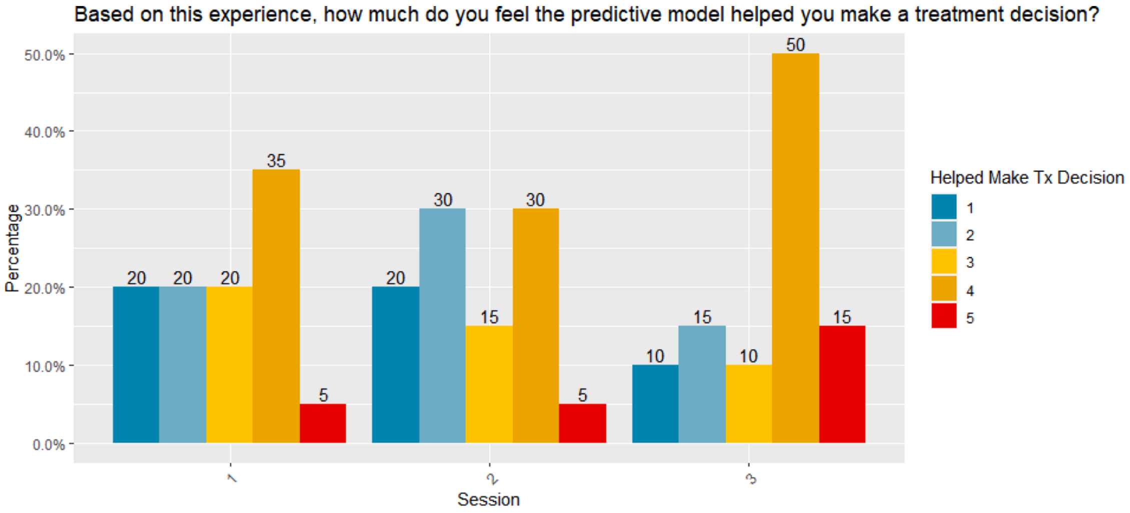

Fig. 4. Ratings of the model's helpfulness in making treatment decisions across simulation sessions (scale of 1 to 5, 1 being "very little" and 5 being "very much").

psychiatrist ratings of model reasonableness.

Taken together, the results of Figs. 1 and 2 suggest that PCPs and psychiatrists both find the model generally produces reasonable predictions, but that PCPs find these predictions more helpful in making treatment decisions. This is discussed below and is summarized in Table 2.

Table 2
Differences Between PCPs and Psychiatrists.

Question	Group	Mean	Standard Deviation
Model Usefulness Rated at Study End	All Participants	3.6	1.1
The predictive model was useful in helping me make treatment decisions (1 = strongly disagree, 5 = strongly agree)	Primary Care Physicians	4.0	1.2
	Psychiatrists	3.2	1.0
Trust in Model by Session			
Based on this experience, how much do you trust the predicted model?			
Session 1	All Participants	3.3	0.9
	Primary Care Physicians	3.3	0.7
	Psychiatrists	3.3	1.1
Session 2	All Participants	3.3	1.0
	Primary Care Physicians	3.3	0.9
	Psychiatrists	3.2	1.1
Session 3	All Participants	3.5	1.1
	Primary Care Physicians	3.6	1.3
	Psychiatrists	3.5	1.0
Model Usefulness by Session			
Based on this experience, how much do you teel the predictive model helped you make a treatment decision?			
Session 1	All Participants	2.9	1.3
	Primary Care Physicians	3.1	1.3
	Psychiatrists	2.6	1.3
Session 2	All Participants	2.7	1.3
	Primary Care Physicians	3.0	1.5
	Psychiatrists	2.5	1.0
Session 3	All Participants	3.5	1.2
	Primary Care Physicians	3.8	1.4
	Psychiatrists	3.2	1.1

Participants were asked to rate their trust in the model on a scale of 1–5 immediately after each clinical experience. From the first session to the third, the proportion of higher ratings of trust increased (Fig. 3). We also note an overall increase in the number of participants responding that they had a somewhat low trust in the model (a 1 or 2 out of 5). Overall, the proportion of participants with some level of trust (a 4 or 5 out of 5) increased from 45% at session 1 to 65% at session 3 (a 20% absolute increase) and the proportion of participants with a negative rating of trust in the model (a 1 or 2 out of 5) also increased, from 15% to 25% (a 10% absolute increase), partially driven by the reduction in participants with a neutral impression as they decided whether to trust the model or not over time. Overall, a small but non-significant increase in trust ratings was found between time 1 (mean 3.3, SD = 0.92) and time 3 (mean 3.5, SD = 1.1) for all participants. For psychiatrists, first session mean trust rating was 3.3, SD 0.9 and third session mean trust rating was 3.5, SD 1.0; for PCPs, first session mean trust rating was 3.3, SD 0.7 and third session mean trust rating was 3.6, SD 1.3.

Participants were also asked how much they felt the predictive model helped them make a treatment decision on a scale of 1–5 immediately after each session. From the first to the third session, there appeared to be an increase in the number of higher ratings on the scale (i.e., greater feeling that the model helped make the treatment decision); for all participants, the first session mean rating was 2.9, SD 1.3 and the third session mean rating was 3.5, SD 1.2. For psychiatrists, first session mean rating was 2.6, SD 1.3 and third session mean rating 3.2, SD 1.1; for PCPs, first session mean rating was 3.1, SD 1.3 and third session mean rating 3.8, SD 1.4 (Fig. 4, Table 2).

In addition, nearly half (48%) of the time, physicians chose one of the top two treatments predicted by the AI. In the supplementary materials, we present an analysis and discussion of predictors and correlates of physician selection of treatments consistent with AI-defined optimal treatments; this includes the result that patient severity and physician trust interact to predict physician selection of treatment options coherent with the AI predictions. Results of the CANMAT quiz can also be found in the supplementary materials.

3.4. Perceived utility — qualitative analysis

Our qualitative analysis provided additional insight into how physicians felt using our tool for the first time and their perceptions of its utility. After reading through the qualitative data, four RAs reduced the original nine themes to the following four to reduce the overlap between excerpts: interpretability of the tool, impact on treatment decision and clinical practice, trust and understanding of AI, and impact on physician-patient interaction. Each of the four themes had a number of sub-themes (Supplementary Table 3 and Supplementary Figure 2). Data presented will focus on the subset pertaining to the perceived utility of the tool. As a general note, RAs observed that use of the tool became progressively more integrated and that the tool was utilized more confidently across trials. Physicians reported that the tool was easy to use, but many nonetheless felt they would have benefited from more practice prior to the clinical scenarios. Moreover, a shared sentiment among several physicians was that in their usual clinical practice, they would likely review some of the information presented in the tool prior to a session with their patient in order to have more time to digest the information.

3.4.1. Trust, understanding, feelings about AI

Participant trust and belief in the AI predictive model was generally positive. Of the four participants who explicitly mentioned “trust” or “faith” with respect to the AI, two reported that they trusted the tool, one wondered about how much to trust the remission probabilities, and one reported having less trust in the AI after seeing one of the important features behind a prediction (which they did not consider to be predictive) but felt that the use of this tool would be especially beneficial in gaining credibility with younger patients during treatment. Another participant reported that the use of AI in psychiatry could be beneficial by introducing some objectivity. Conversely, one participant worried about the integration of AI models in healthcare, “AI interprets data, but people are not data”.

Furthermore, six participants requested more evidence behind the AI model, specifically the datasets informing the remission probabilities. Of those, two expressed that if more information about the algorithmic reasoning behind treatment decisions had been available, they would have trusted the model more.

Four participants expressed they had a limited understanding of how the AI predictive model functioned, but were keen to integrate it into their treatment decisions. The comfort level of the participant with the AI model, reported via RA observations, was aligned with how well they were able to explain the model to the SP. The terminology used to describe the tool to the SPs varied greatly, including “tool”, “new technology”, “the algorithm”, and “the computer”. Interestingly, only 25% of participants used the term “AI” to describe the model, with another 25% preferring to describe it, for example, as a tool that “makes decisions based on data from many other individuals”.

3.4.2. Impact on treatment decision and clinical practice

Participants were asked about the impact of the model on their treatment decision. Roughly half (45%) of the responses were distinctly positive, describing the potential of the tool to transform practice or diversify treatment approaches. One participant noted that the CDSS “helps you to either choose a specific antidepressant medication or at least narrow down your range of choices to a few ideal candidates”. 25% of participants reported no impact on treatment decisions from the tool due to greater confidence in their own clinical judgment, perceived minimal differences in projected outcomes between suggested treatments, and/or noting that treatments suggested by the model were already in line with their prior treatment plan. Negative comments centered around the interference of the tool in the physician-patient relationship, or, for one participant, on the perception that the tool focused too greatly on medication.

A recurring theme was related to the significance of remission rate percentages offered by the tool: if the probabilities were felt to be clinically significant, participants described a greater impact on their treatment decision, “normally [I] wouldn’t prescribe a medication for [this] case [… but I will] because the percentage is quite high so I do think that it is worth trying the medication”; this is juxtaposed with the perception of little to no impact on treatment decision if the differences between percentages was interpreted as insignificant “[a] small difference in percentages is not going to change how I practice”.

Participants described their perceived potential value of the tool in several ways: as a potential way to save time (10%); to confirm/suggest treatment options (15%); as a centralized tool for guidelines (10%); as a source of extra information about treatments (15%); as useful for displaying symptoms over time (5%, though it should be noted that a longitudinal data element was not included in this study); and as a way to explain to patients why a particular treatment is being chosen (15%). The value of gaining more familiarity with the model was also apparent from participant comments: “since this was the first time using it I did tend to stick to what I would usually prescribe, however I can see that if I got used to using it regularly as part of my residency training that I would probably use different treatment options”.

3.4.3. Communicability and interpretability of the CDSS’s results

The key to the clinical utility of a tool aimed at supporting shared decision-making is the ability of physicians to communicate results and their impact on decision-making to patients, and, relatedly, physician understanding of the tool. 40% of physicians made reference to the benefit of being able to communicate the motivation for a treatment decision to the patient, “especially to give patients concrete numbers about their remission probabilities”. As one physician observed, “One good thing was you could explain to the patient why you are choosing the treatments that you are choosing”. 20% of participants expressed wanting more information about the source of the remission probabilities, either about the model itself, the clinical data that trained the deep learning model, or about how the variables considered by the model impacted the remission prediction for an individual patient.

4. Discussion

In order to support clinical decision-making, a CDSS must provide high-quality, clinically useful information that is relevant to the individual patient while being accessible, interpretable, and actionable for the clinician (Sim et al., 2001). Additionally, the effort expended by the physician to use the system must not be perceived as excessive (Wendt et al., 2000). In this study of an AI-powered CDSS for depression treatment with a sample of intended users (primary care and psychiatry staff and residents), we aimed to evaluate the perceived utility of the tool, the perceived impact of the tool on clinical decision-making, and potential differences in the perceived utility between PCPs and psychiatrists. In the supplementary materials, we also discuss the drivers of physician prescription of treatments consistent with those predicted by the AI model as having the highest likelihood of success. Conducting this study using simulated patients created a safe and controlled environment in which to measure the tool’s impact on the treatment decision-making process (Benrimoh et al., 2021).

In a previous paper reporting on this dataset (Benrimoh et al., 2021), we demonstrated that physicians felt the tool was feasible to use in a clinical interaction and did not have significantly deleterious effects on the patient-clinician interaction. Once a CDSS has met the basic requirement of being easy to use within a reasonable time frame (in our study, in roughly five minutes within a ten-minute interview), can be conceivably worked into the clinical workflow, and seems to garner the trust of most users; the next question to be answered is therefore: do physicians find the tool to be useful with respect to its intended goal of assisting in treatment decisions? We informed our participating physicians that they were free to use or ignore the CDSS predictions and to choose treatments as they saw fit. This provided us with an opportunity to use a mixed-methods approach to investigate their perceived utility of the system in a controlled environment, where their interactions with the CDSS and with patients could be observed without being directed (i.e. they were not forced to use or pay attention to all or part of the tool).

Overall, the perceived usefulness of the model in making treatment decisions increased from session 1 to session 3.In line with the qualitative results, this indicates that improved familiarity with the CDSS and the resulting increased comfort allowed clinicians to better understand and integrate it into their approach, improving its perceived utility. In the first session, 40% of participants felt the model was useful in making treatment decisions (rating a 4 or 5 out of 5), and this increased to 65% by the third session. In contrast, 40% of participants did not feel the model was useful in the first session (rating 1 or 2 out of 5), and this number decreased to 25% by the third session. With respect to trust, a large proportion of physicians (40%) began with neutral feelings regarding their trust towards the model, which is reasonable as they had not used it with a patient at that point. As physicians decided whether to trust the model or not as they used it, this proportion decreased to 10% by session 3. As time went on, across all physicians, relatively more physicians chose to place some level of trust in the model (with this proportion rising from 45% at session 1 to 65% at session 3; a 20% absolute increase)) than those who decided to trust it less (with this proportion rising from 15% to 25%; a 10% absolute increase). Overall, while these changes are small in the context of our sample size, they do suggest that familiarity with the tool may play an important role in influencing physician trust. In addition, given that some clinicians will have a reduction in trust as they begin using the model, when implementing these systems in practice, continued support should be provided and physicians should be encouraged to raise concerns so that these can inform further improvements to the model and continue building trust.

Feedback from the participating physicians revealed differences between PCPs and psychiatrists. Specifically, PCPs seem to find the model more useful in helping them make treatment decisions (Fig. 1). PCPs have been found to perceive the treatment of patients with depression as challenging, feeling that these patients place a high demand on their psychological resources (McPherson and Armstrong, 2012). This sense that depressed patients present a treatment challenge, combined with the fact that PCPs may not be as familiar with the guidelines or the available range of medications as specialists, may have driven our finding that PCPs are more likely to find the model useful in making treatment decisions. It is interesting to note that psychiatrists and PCPs essentially did not differ in their estimation of the reasonableness of the model's predictions (Fig. 2), or in their trust of the model over sessions (Table 2), indicating that this difference in perceived utility was likely not due to a perceived difference in the validity of model predictions between physician types that might have been influenced by the experience gap between PCPs and psychiatrists, or by their clinical experience of differing patient populations.

The mixed-methods approach of this study yielded qualitative data that can further nuance our understanding of perceived utility. The impact of the CDSS on treatment decision-making was, overall, viewed positively by participants, as evidenced both by the qualitative comments and by the finding at session 3 that 65% of participants felt that the tool was useful in making treatment decisions. Some participants noted the availability of questionnaire data and the predicted remission probabilities as being useful in shared decisionmaking with patients, with some physicians specifically noting that they would use the remission probabilities to help patients understand possible treatment choices. Other elements of potential utility were noted, with some responses suggesting value from potential time savings as well as the ability to centralize information about treatments which can assist in the review of different options available. The ability to use the CDSS to "narrow down" the list of possible treatments was also highlighted. Taken together with the willingness to consider CDSS predictions demonstrated by participants (for example, in the quote described above where a participant physician was willing to consider prescribing a medication when they normally would not), and the fact that treatments chosen agreed with the model's top two choices 48% of the time, this suggests that clinicians see the tool as being a useful aid to decision, but not as a replacement for their clinical judgement, especially when the predicted differences between treatments were small. Indeed, a number of participants chose to favor their own judgment over that of the CDSS, citing their greater confidence in their own experience. These results are in line with the design philosophy behind the CDSS, in that the tool is envisioned as an aid to clinicians and patients in shared decision-making and in facilitating the use of measurement-based, algorithm-guided care, but not as a tool meant to 'hijack' clinical decision-making.

One key result apparent in the qualitative data, which was reflected in the quantitative results (as evidenced by the increased in perceived usefulness of and, for some participants, trust in the model from session 1 to session 3), is the feeling on the part of physicians that greater time to work with and be exposed to the application, as well as an improved understanding of the AI, would improve the ease of use and comfort with the tool, increase trust in the tool, and likely increase its perceived utility for assisting clinical decision-making. For example, as discussed above, one participant physician mentioned they chose to select the treatment they would usually prescribe during the session, but felt that with continued use of the tool they could see it helping them expand the treatment options they would consider. This result supports the finding in the quantitative data that the perceived utility of the model increased over the course of the three sessions. This information is critical because it will influence the design of training materials and procedures, ensuring that longer training periods as well as more comprehensive training materials are provided. In addition, in future clinical studies, the prediction that greater familiarity with the tool will yield increased trust in or utilization of CDSS treatment predictions could be tested.

Limitations

Our findings should be interpreted considering several limitations. The small sample size and a lack of specific endpoints represent limitations for this study, which was underpowered to detect differences between time points. This remains a limitation, despite the objective of the study of exploring participant reactions to this new technology. As such, future work using simulation centers to evaluate perceived utility of these technologies should include a defined primary endpoint, and defined secondary endpoints, and should be appropriately powered to detect them. The small sample size also limits the generalizability of the results and may bias the quantitative analysis (Moineddin et al., 2007), and also prevented the use of extensive testing for significance because of lack of power. As such, trends observed here should act as results to be replicated in the future with larger clinical samples. In addition, the lack of a control group makes it impossible to know how physicians would have acted without access to the CDSS. However, the main purpose of the current study was to establish the ease of use and perceived utility of our CDSS. Future work in clinical populations will assess the feasibility, safety, clinical utility, and effectiveness of the CDSS and seek to replicate the current results in larger clinical samples. Given the limitations on time in the simulation center, participants were provided with only brief training without the opportunity for repeated practice sessions. We observed that participant scores that rated dimensions of trust, ease, and helpfulness increased on average from the first trial to the last trial, suggesting that, as expected, comfort level with the CDSS improves as a function of practice. Moreover, participants expressed an interest in more extensive training and practice sessions, as well as more extensive information about the AI model. These will be provided in future clinical studies alongside extended training, but the limited training in this study may have impacted perceived utility. Finally, a simulation center is not a real clinical setting and testing in real clinical environments will be required to replicate and verify clinician perceived utility and verify the validity of these results in usual clinical environments. A central question remains whether clinicians will use and continue to use this tool in real clinical practice, or if there will be unexpected barriers to its use or drop off of its use with time. These questions are beyond the scope of what can be addressed in a simulation setting and require further clinical research.

Conclusion

We present preliminary findings of perceived clinical utility of a novel AI-enabled CDSS for physicians treating patients with depression. Overall, physicians found the system useful and beneficial during shared decision-making with patients, with PCPs perceiving the greatest utility. Physicians perceived the tool to be more useful with repeated use, consistent with their comments that greater training and time with the tool would likely increase their perceived utility. This CDSS presents a new opportunity to use readily available patient data to personalize treatment choice at the point of care, and preliminary results indicate that physicians are able and willing to use this kind of tool to support their decision-making. Further advances in AI interpretability as well as improved training regimens for physicians should help improve trust and, in turn, use of AI results. Establishment of this kind of tool in the treatment of depression may lead to applications in other areas of mental health. The use of a mixed methods approach as well as the simulation center was useful in providing information that could benefit further development of the CDSS and improve training for participants in future studies. Clinical trials are nonetheless required to assess the effectiveness of this tool in improving mental health outcomes. These trials will help determine the utility of the CDSS from the patient's perspective, which is necessary to build a tool that delivers care aligned with patients' needs and preferences. Creating a CDSS via a patient-centric approach has the potential to improve the support provided to patients and empower them to participate more in their own care.

Funding sources

Aifred Health Inc.; Innovation Research Assistance Program, National Research Council Canada; ERA-Permed Vision 2020 supporting IMADAPT; Government of Quebec Nova Science; MEDTEQ COVID-Relief Grant. Use of the simulation center and the work of the standardized patients was provided as part of the prize for a clinical innovation competition run by McGill University and the Steinberg center for Simulation and Interactive Learning, Canada, with the generous support of the Hakim family.

CRediT authorship contribution statement

Myriam Tanguay-Sela: Methodology, Resources, Investigation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing, Supervision, Project administration. **David Benrimoh:** Conceptualization, Methodology, Software, Resources, Investigation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing, Supervision. **Christina Popescu:** Formal analysis, Investigation, Visualization, Supervision, Writing – original draft, Writing – review & editing. **Tamara Perez:** Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Colleen Rollins:** Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Emily Snook:** Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Eryn Lundrigan:** Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Caitrin Armstrong:** Software, Validation, Writing – review & editing. **Kelly Perlman:** Conceptualization, Resources, Investigation, Writing – review & editing. **Robert Fratila:** Software, Validation, Writing – review & editing. **Joseph Mehltretter:** Software, Validation, Writing – review & editing. **Sonia Israel:** Conceptualization, Resources, Investigation, Writing – review & editing. **Monique Champagne:** Writing – review & editing. **Jérôme Williams:** Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Jade Simard:** Writing – review & editing. **Sagar V. Parikh:** Conceptualization, Writing – review & editing. **Jordan F. Karp:** Conceptualization, Writing – review & editing. **Katherine Heller:** Conceptualization, Writing – review & editing. **Outi Linnaranta:** Conceptualization, Writing – review & editing. **Liliana Gomez Cardona:** Conceptualization, Writing – review & editing. **Gustavo Turecki:** Conceptualization, Writing – review & editing. **Howard C. Margolese:** Conceptualization, Writing – review & editing.

Conflicts of interest

M.T.S. is employed by Aifred Health and is an options holder. D.B., C. A., K.P., R.F., and S.I. are shareholders and either employees, directors, or founders of Aifred Health. C.P., T.P., C.R., E.S., E.L., J.M., M.C., J.S., and J.W. have been or are employed or financially compensated by Aifred Health. S.V.P., J.F.K., and K.H. are members of Aifred Health's scientific advisory board and have received payments or options. H.C.M. has received honoraria, sponsorship or grants for participation in speaker bureaus, consultation, advisory board meetings and clinical research from Acadia, Amgen, HLS Therapeutics, Janssen-Ortho, Mylan, Otsuka-Lundbeck, Perdue, Pfizer, Shire and SyneuRx International. All other authors report no relevant conflicts.

Acknowledgements

We would like to acknowledge the Steinberg Centre for Simulation and Interactive Learning for the helpfulness of their staff in assisting with the execution of this study, as well as the standardized patients who participated in the study, for their excellence and the quality of their feedback.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2021.114336.

References

Archibald, M.M., 2016. Investigator triangulation: a collaborative strategy with potential for mixed methods research. J Mix Methods Res 10, 228–250. https://doi.org/10.1177/1558689815570092.

Benrimoh, D., Israel, S., Perlman, K., Fratila, R., Krause, M., 2018. Meticulous transparency—an evaluation process for an agile ai regulatory scheme. In: Mouhoub, M., Sadaoui, S., Ait Mohamed, O., Ali, M. (Eds.), Recent Trends and Future Technology in Applied Intelligence, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 869–880. https://doi.org/10.1007/978-3-319-92058-0_83.

Benrimoh, D., Tanguay-Sela, M., Perlman, K., Israel, S., Mehltretter, J., Armstrong, C., Fratila, R., Parikh, S.V., Karp, J.F., Heller, K., Vahia, I.V., Blumberger, D.M., Karama, S., Vigod, S., Myhr, G., Martins, R., Rollins, C., Popescu, C., Lundrigan, E., Snook, E., Wakid, M., Williams, J., Soufi, G., Perez, T., Tunteng, J.-F., Rosenfeld, K., Miresco, M., Turecki, G., Cardona, L.G., Linnaranta, O., Margolese, H.C., 2021. 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, 1–10. https://doi.org/10.1192/bjo.2020.127.

Beullens, J., Rethans, J.J., Goedhuys, J., Buntinx, F., 1997. The use of standardized patients in research in general practice. Fam. Pract. 14, 58–62. https://doi.org/10.1093/fampra/14.1.58.

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qual. Res. Psychol. 3, 77–101. https://doi.org/10.1191/1478088706qp063oa.

Bromet, E., Andrade, L.H., Hwang, I., Sampson, N.A., Alonso, J., de Girolamo, G., de Graaf, R., Demyttenaere, K., Hu, C., Iwata, N., Karam, A.N., Kaur, J., Kostyuchenko, S., Lépine, J.-P., Levinson, D., Matschinger, H., Mora, M.E., Browne, M.O., Posada-Villa, J., Viana, M.C., Williams, D.R., Kessler, R.C., 2011. Cross-national epidemiology of DSM-IV major depressive episode. BMC Med. 9, 1–16. https://doi.org/10.1186/1741-7015-9-90.

Cheng, A., Kessler, D., Mackinnon, R., Chang, T.P., Nadkarni, V.M., Hunt, E.A., Duval-Arnould, J., Lin, Y., Cook, D.A., Pusic, M., Hui, J., Moher, D., Egger, M., Auerbach, M., International Network for Simulation-based Pediatric Innovation, Research, and Education (INSPIRE) Reporting Guidelines Investigators, 2016. Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. Simul. Healthc. 11, 238–248. https://doi.org/10.1097/SIH.0000000000000150.

Hamilton, M., 1960. A rating scale for depression. J. Neurol. Neurosurg. Psychiatry 23, 56–62.

Harrison, P., Carr, E., Goldsmith, K., Young, A.H., Ashworth, M., Fennema, D., Barrett, B., Zahn, R., 2020. Study protocol for the antidepressant advisor (ADESS): a decision support system for antidepressant treatment for depression in UK primary care: a feasibility study. BMJ Open 10, 1–9. https://doi.org/10.1136/bmjopen-2019-035905.

Kennedy, S.H., Lam, R.W., McIntyre, R.S., Tourjman, S.V., Bhat, V., Blier, P., Hasnain, M., Jollant, F., Levitt, A.J., MacQueen, G.M., McInerney, S.J., McIntosh, D., Milev, R.V., Müller, D.J., Parikh, S.V., Pearson, N.L., Ravindran, A.V., Uher, R., 2016. 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. Psychiat. 61, 540–560. https://doi.org/10.1177/0706743716659417.

Kroenke, K.L., Spitzer, R.L., Williams, J.B., 2001. The PHQ-9: validity of a brief depression severity measure. J. Gen. Intern. Med. 16, 606–613.

McPherson, S., Armstrong, D., 2012. General practitioner management of depression: a systematic review. Qual. Health Res. 22, 1150–1159. https://doi.org/10.1177/1049732312448540.

Mehltretter, J., Fratila, R., Benrimoh, D., Kapelner, A., Perlman, K., Snook, E., Israel, S., Armstrong, C., Miresco, M., Turecki, G., 2020. Differential treatment benefit prediction for treatment selection in depression: a deep learning analysis of STAR*D and CO-MED data. Computat. Psych. 4, 61–75. https://doi.org/10.1162/cpsy_a_00029.

Moineddin, R., Matheson, F.I., Glazier, R.H., 2007. A simulation study of sample size for multilevel logistic regression models. BMC Med. Res. Methodol. 7, 34. https://doi.org/10.1186/1471-2288-7-34.

Rollman, B.L., Hanusa, B.H., Lowe, H.J., Gilbert, T., Kapoor, W.N., Schulberg, H.C., 2002. A rando- mized trial using computerized decision support to improve treatment of major depression in primary care. J. Gen. Intern. Med. 17, 493–503.

Rush, A.J., Carmody, T., Reimitz, P.-.E., 2000. The inventory of depressive symptomatology (IDS): clinician (IDS-C) and self-report (IDS-SR) ratings of depressive symptoms. International Journal of Methods in Psychiatric Research, 9 (2), 45–59. https://doi.org/10.1002/mpr.79.

Rush, A.J., Trivedi, M.H., Ibrahim, H.M., Carmody, T.J., Arnow, B., Klein, D.N., Markowitz, J.C., Ninan, P.T., Kornstein, S., Manber, R., Thase, M.E., Kocsis, J.H., Keller, M.B., 2003. The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol. Psychiatry 54, 573–583. https://doi.org/10.1016/s0006-3223(02)01866-8.

Shirazi, M., Sadeghi, M., Emami, A., Kashani, A.S., Parikh, S., Alaeddini, F., Arbabi, M., Wahlstrom, R., 2011. Training and validation of standardized patients for unannounced assessment of physicians' management of depression. Acad. Psychiatry 35, 382–387. https://doi.org/10.1176/appi.ap.35.6.382.

Sim, I., Gorman, P., Greenes, R.A., Haynes, R.B., Kaplan, B., Lehmann, H., Tang, P.C., 2001. Clinical decision support systems for the practice of evidence-based medicine. J. Am. Med. Inform. Assoc. 8, 527–534. https://doi.org/10.1136/jamia.2001.0080527.

Sutton, R.T., Pincock, D., Baumgart, D.C., Sadowski, D.C., Fedorak, R.N., Kroeker, K.I., 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. npj Digital Med. 3, 1–10. https://doi.org/10.1038/s41746-020-0221-y.

Trivedi, M.H., Kern, J.K., Grannemann, B.D., Altshuler, K.Z., Sunderajan, P., 2004. A computerized clinical decision support system as a means of implementing depression guidelines. Psychiat. Serv. 55, 879–885.

Warden, D., Rush, A.J., Trivedi, M.H., Fava, M., Wisniewski, S.R., 2007. The STAR*D project results: a comprehensive review of findings. Curr. Psychiat. Rep. 9, 449–459.

Wendt, T., Knaup-Gregori, P., Winter, A., 2000. Decision support in medicine: a survey of problems of user acceptance. Stud. Heal. Technol. Inform. 77, 852–856.

World Health Organization, 2017. Depression and other common mental disorders: global health estimates. https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf (accessed 10 January 2021).

Zikos, D., DeLellis, N., 2018. CDSS-RM: a clinical decision support system reference model. BMC Med. Res. Methodol. 18, 137. https://doi.org/10.1186/s12874-018-0587-6.