CrossMark

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

David Benrimoh*, Myriam Tanguay-Sela*, Kelly Perlman, Sonia Israel, Joseph Mehltretter, Caitrin Armstrong, Robert Fratila, Sagar V. Parikh, Jordan F. Karp, Katherine Heller, Ipsit V. Vahia, Daniel M. Blumberger, Sherif Karama, Simone N. Vigod, Gail Myhr, Ruben Martins, Colleen Rollins, Christina Popescu, Eryn Lundrigan, Emily Snook, Marina Wakid, Jérôme Williams, Ghassen Soufi, Tamara Perez, Jingla-Fri Tunteng, Katherine Rosenfeld, Marc Miresco, Gustavo Turecki, Liliana Gomez Cardona, Outi Linnaranta and Howard C. Margolese

Background

Recently, artificial intelligence-powered devices have been put forward as potentially powerful tools for the improvement of mental healthcare. An important question is how these devices impact the physician-patient interaction.

Aims

Aifred is an artificial intelligence-powered clinical decision support system (CDSS) for the treatment of major depression. Here, we explore the use of a simulation centre environment in evaluating the usability of Aifred, particularly its impact on the physician–patient interaction.

Method

Twenty psychiatry and family medicine attending staff and residents were recruited to complete a 2.5-h study at a clinical interaction simulation centre with standardised patients. Each physician had the option of using the CDSS to inform their treatment choice in three 10-min clinical scenarios with standardised patients portraying mild, moderate and severe episodes of major depression. Feasibility and acceptability data were collected through self-report questionnaires, scenario observations, interviews and standardised patient feedback.

Results

All 20 participants completed the study. Initial results indicate that the tool was acceptable to clinicians and feasible for use during clinical encounters. Clinicians indicated a willingness to use the tool in real clinical practice, a significant degree of trust in the system's predictions to assist with treatment selection, and reported that the tool helped increase patient understanding of and trust in treatment. The simulation environment allowed for the evaluation of the tool's impact on the physician–patient interaction.

Conclusions

The simulation centre allowed for direct observations of clinician use and impact of the tool on the clinician–patient interaction before clinical studies. It may therefore offer a useful and important environment in the early testing of new technological tools. The present results will inform further tool development and clinician training materials.

Keywords

Primary care; out-patient treatment; depressive disorders; artificial intelligence; simulation centre.

Increasingly, new technologies that supplement clinical decision-making are being implemented to respond to the need to improve mental health treatment outcomes.[1] Some of these tools are designed to be used at the point of care, during sessions with patients, and may be expected to have some impact on physician–patient interactions, which may, in turn, affect the physician–patient relationship, one of the most critical aspects of psychiatric intervention.[2] It is challenging to test the effect of tools on these interactions and on clinical workflow, as directly observing clinical interviews can be impractical or raise concerns about the validity of observations.

We assess the use of simulation to directly observe the impact of an artificial intelligence-powered decision support tool on simulated patient–clinician interactions. The objective was to determine if and how the use of the tool during a session affected on the physician–patient interaction, as a prelude to longitudinal clinical studies assessing longer-term effects on clinical workflow and the physician–patient relationship. Using simulation, clinician behaviour can be observed in a secure setting,[3] and data can be collected from multiple viewpoints, i.e. that of the clinician, the standardised patient and the observer. This triangulation process is a rigorous method for gathering high-quality data.[4] We discuss the challenges encountered and insights gained from our experience with simulation-based testing of new technology.

Background on depression treatment challenges

As noted, in this paper we focus on the simulation centre testing of a tool aimed at supporting clinical decision-making during treatment selection for depression. This is an important field of work because depression is a common condition, with over one in nine people

---

* These authors contributed equally to this work.

experiencing it in their lifetime,[5] with a high burden, now being the leading cause of disability globally.[6,7] Although many people with depression remain undiagnosed,[8] among those who are and who receive treatment only roughly a third will achieve remission during a first treatment course,[9] with many patients needing to go through multiple treatment trials before finding an effective treatment. Physicians (both psychiatrists and primary care physicians) are faced with a large selection of effective treatments, as well as guidelines which help to manage treatments once they are chosen, but they do not currently have access to tools that can help them effectively choose between the existing first-line agents to optimise chances of treatment success and minimise the need for repeated trial-and-error treatment trials.[10] This need for improved decision support has led to a number of projects aimed at improving the personalisation of treatment selection, notably pharmacogenomics.[11] However, pharmacogenomics may be expensive, and samples may take time to be processed, which could be used to treat the patient. One solution would be a tool that can assist with the personalisation of treatment at the point of care, using readily available clinical and demographic data; for this purpose, a number of researchers[12,13] have explored the use of machine learning and artificial intelligence as collections of techniques that can assess complex patterns (such as are found in patient data) and link them to outcomes (such as remission). In this paper, the tool discussed utilises artificial intelligence to provide clinicians with estimates of the likely efficacy of different treatments, to assist them in shared decision-making about which agent to try first with their patient. Future iterations of the tool will extend this to decision-making after treatment failure. Regardless of the specific point in the care pathway of a given patient, learning about the acceptability and useability of these kinds of clinical decision support systems (CDSSs) will be key to maximising their clinical impact, and this was a key purpose of the present study. This study was meant to observe how clinicians interact with the tool, as a step in its development and the development of training protocols for clinical studies involving the tool.

Aifred: clinical decision support software for depression treatment

We investigated the use of Aifred, a CDSS that includes an operationalised version of the 2016 Canadian Network for Mood and Anxiety Treatments (CANMAT) guidelines for depression treatment,[14] and provides artificial intelligence decision support when treatments are chosen. This artificial intelligence helps support clinicians by considering complex interactions between multiple patient variables to help personalise treatment in order to improve upon a trial-and-error treatment approach and reduce the number of failed treatment trials.[10,15] It also tracks symptoms by using standardised questionnaires such as the Patient Health Questionnaire-9.[16] Major depressive disorder (MDD) was chosen, given its high prevalence,[17,18] status as the leading cause of disability globally[19] and poor remission rates following initial treatment.[9]

The key innovation is the inclusion of an artificial intelligence tool that provides clinicians with remission probabilities for different treatment options, based on a patient's clinical and demographic profile. This artificial intelligence is layered on top of the operationalised CANMAT guidelines, providing remission probabilities for individual treatments at the point in the guideline when the first-line treatment is chosen. The expected clinical utility of this artificial intelligence model is as follows. As noted, clinicians currently mostly follow a trial-and-error pattern when selecting treatments for depression, and, beyond providing a pool of first-line treatments, the guidelines are not able to precisely guide the selection of individual agents at the beginning of treatment. Although at

the population level these treatments are considered to be essentially equally effective,[20] the low rates of remission after initial treatment,[9] the varying pharmacological profiles and different efficacy of even similar antidepressants,[14] and the clinical observation that different patients seem to respond to different treatments, have resulted in efforts to try and identify patterns with machine-learning tools, so as to predict the efficacy of specific agents for individual patients based on clinical and demographic information,[12] or combining this information with biomarkers.[21] The Aifred tool provides remission probabilities for a number of treatments simultaneously, providing the clinician with extra information to help select a treatment within the pool of those recommended by the guidelines. This is meant to provide an estimate of likely treatment benefit, which can be used, alongside consideration of side-effects, medical history and patient preferences, with the intention of optimising treatment choice and reducing the chance a patient will start a treatment that is less likely to help them reach remission. Without these probabilities, there is very little information available to help clinicians select between first-line treatments with respect to their likely efficacy. This artificial intelligence tool is a deep-learning model, trained and validated on baseline clinical and demographic data from 4735 patients from five major studies[15] (STAR*D[9], CO-MED[22], EMBARC[23], REVAMP[24] and IRL-GREY[25]). Patient clinical and demographic features, such as fatigue, physical symptoms and employment status, were identified with a feature selection pipeline described in Mehltretter et al,[15] and were then used to train a deep neural network. This network's objective was to predict patient remission status, and the drug assigned to the patient in the study was retained as a predictive feature. Once the model was trained, probabilities for remission for each treatment for a new patient could be derived by feeding that patient's clinical and demographic data into the model and then iterating over each of the possible treatments via the treatment-assigned variable. The model currently provides individualised remission probabilities for five commonly used first-line treatments (escitalopram, citalopram, bupropion, venlafaxine and sertraline) and two combination treatments (bupropion plus escitalopram, and venlafaxine plus mirtazapine). The remission probabilities are presented as follows: for each treatment for which a probability can be calculated, a raw remission probability (e.g. 45%) is presented next to the name of the treatment. This probability represents the chance that the individual patient in question will reach remission, assuming appropriate use of the treatment as per guidelines, and an appropriate treatment trial. By clicking on a button labelled 'more', included next to each treatment, clinicians were able to see the baseline population remission rate based on the data-set used to train the model (in our case, this was 34.85%), as well as the 'interpretability report'. This report was a list of up to five of the patient variables that were most important in producing the probability for that drug for the given patient; these were derived using a feature importance algorithm described in Mehltretter et al,[15] which would produce different sets of features for each treatment, for each patient. *In silico* testing of this model demonstrated that it is potentially capable of improving population remission rates (testing methods described in Mehltretter et al[15]). Future versions of this model are planned to increase the number of predicted treatments, and also to include psychotherapies and augmentation treatments. Note, however, that the focus of this paper is not on the specific artificial intelligence model (which may continue to evolve until the start of clinical trials), but of the impact of such a model, packaged within a digital health platform, on the patient–clinician interaction.

It should also be noted that the integrated CANMAT guidelines provide the most support in terms of the longitudinal management of depression treatment (i.e. when to switch or augment treatments

in the case of poor response), and as such in this study, which focused on a singular interaction, functioned mostly to provide an evidence-based pool of initial treatment options that could be differentiated by the artificial intelligence model on a patient-by-patient basis, as well as guideline-derived treatment initiation advice (for example, by reminding clinicians of the benefit of combining pharmacotherapy and psychotherapy). In future studies, the combined effect of longitudinal management using the guidelines and the optimisation of initial treatment selection using the artificial intelligence will be studied, but is out of scope for the present paper.

The tool is intended to be used during patient interviews, providing access to evidence-based decision support. It was designed with a simple interface intended to minimise time spent clicking through menus so that clinicians could focus on reviewing data and the artificial intelligence results, ideally while discussing and viewing them with their patient as part of shared decision-making. Numerical remission probabilities are provided for those treatments on which the model is trained, but clinicians can choose from any of the treatments appearing in CANMAT. This simulation study sought to assess whether the tool, which should always be employed in the context of best clinical judgement and patient preference, could be feasibly used at the point of care as well as maintaining, or possibly enriching, the integrity of the physician–patient interaction.

When designing the Aifred tool, one of the primary considerations was how the tool could support shared decision-making between clinicians and patients, in accordance with best practices.[26] Indeed, the tool was developed using an informal participatory process where patient input was sought on design during development, and several members of the core development team had lived experience with depression and other mental health conditions, and had experienced treatment selection interactions with clinicians. The tool also at a number of points makes reference to the importance of discussing treatment preferences with patients, as per best practices.[26] However, despite the fact that shared decision-making is an integral part of good clinical practice, the fact remains that not all clinicians engage in shared decision-making at all times,[26] and the format of this may change in a clinician-dependent manner. In the context of the deployment of a new tool, we decided to observe how clinicians interact with this tool and use, or not use, it as part of shared decision-making without being explicitly prompted on how to do so. This is why the computer was chosen to be a laptop (which can be easily moved) and why it was positioned at 45 degrees (i.e. with the screen part-way between the patient and the clinician, to allow it to be moved one way or the other and remain in a comfortable position for the clinician to begin using). This provided a useful setup to observe clinician behaviour (i.e. to see if they would turn the screen toward the patient or turn it toward themselves, potentially even before they have had a chance to read prompts on the screen), and then to get feedback from standardised patients about how different clinician approaches to using the tool affected their experience.

Previous decision support research

Although previous studies have suggested that treatment utilising a clinical decision algorithm and measurement-based care lead to better patient outcomes,[27,28] often these studies included support from a clinical team or other non-computerised support.[27,29] As such, it is worth reviewing previous work aimed at using computer-based CDSSs to improve depression treatment. Rollman et al[30] created a system that helped screen patients for depression and then offered guideline-based treatment advice messages. In a study of 200 patients in primary care, this tool did not show a positive effect on patient outcomes at 3 or 6 months. One major

technical limitation of this system was that the tool relied on research assistants to program advice messages, and these were not sent to the clinician during clinical encounters, which may have limited its utility. The Texas Medication Algorithm Project (TMAP) led to the development of a computerised version of its clinical algorithm, called CompTMAP, which assisted physicians in decisions such as adjusting doses, starting augmentation treatments and following patient progress in an expert guideline-informed manner.[31] This tool was tested in an unblinded study of 55 patients, where the group of patients treated using the CDSS showed improvement over standard of care in terms of patient depression symptoms.[32] More recently, Harrison et al[33] published a protocol for an upcoming study of a computerised decision support system implementing National Institute of Health and Care Excellence guidelines, which appears to, similarly to CompTMAP, take in patient information and suggest treatment approaches depending on treatment response and the relevant sections of the guidelines. Although all three of these systems offer the ability to screen patients, follow their response to treatment and suggest treatment course changes based on patient response and relevant guidelines (i.e. they support treatment management), none offer the ability to personalise treatment choice and differentiate between specific treatments based on an individual patient's profile (beyond making suggestions about when to alter treatment or add an augmenting agent, as per criteria set out by the guidelines). In the study of depressed inpatients carried out by Adli et al,[28] one arm of the study included a computerised system that did have some extent of prediction based on individual patient data: it used data from 650 patients to calculate probabilities of treatment failure or success during follow-up based on depression symptom scores for an individual patient, although it only provided general advice in response to this. For example, the authors state that the system could provide a recommendation that a physician review the treatment or consider an augmenting agent; as such, this system performed in a similar fashion to the guidelines (which already recommend treatment changes based on clinical improvement, or lack thereof, based on symptom scores at different points in treatment) and was outperformed by a more specific, structured clinical treatment algorithm. As such, no system before Aifred, to the best of our knowledge, combines the ability to implement clinical practice guidelines during patient encounters and patient follow-up (that is, optimising treatment management) with a machine-learning system that provides patient-and-drug specific remission probabilities (i.e. with a view to optimising personalised treatment selection). In this study, we focused on the most novel component offered by Aifred – this personalisation component – to determine how its integration into the information available to a clinician during a patient interaction, using a computerised CDSS, might affect the patient–clinician interaction, with a view to using this information to inform the conduct of future studies of this tool.

Method

For the present study, the sample consisted of the intended end-users of the CDSS: psychiatry and family medicine attending staff and residents. Participants were recruited via email, social media and announcements, and were compensated. The recruitment target was 25 participants. Recruitment started roughly 3 months before study start. This study was approved by the Research Ethics Board of the Douglas Mental Health University Institute (ethical approval number: IUSMD 18-03). All participants, including standardised patients, provided written informed consent to participate. The study was conducted in accordance with the Tri-Council Statement on research ethics.

The study was conducted at the Steinberg Centre for Simulation and Interactive Learning. Each participant was present at the simulation centre for one 2.5-h session. The centre's one-way mirror system allowed research assistants to observe scenarios. The simulation centre has a roster of professional actors who play standardised patients (SPs). The ability of SPs to standardise their acting[34,35] allows for multiple equivalent instances of the same clinical scenario to be run. Research assistants wrote observations on data extraction forms created for the study.

We created three clinical situations, corresponding to a mild, moderate and severe MDD. These situations were based on data from real patients drawn from the de-identified data-sets on which the model was trained. 'Jack' was a retired White male in his 80 s, suffering a mild depression marked by social withdrawal and sleep disturbance. He was experiencing some guilt about a previous divorce. 'Emma' was a White professional female in her 40 s, suffering from moderate depression marked by agitation and guilt about poor performance at work and with respect to being emotionally unavailable within her couple. 'Sara' was an Black female in her 50 s who had lost her job because of severe depression marked by psychomotor retardation and fatigue. She was prompted to come in to see the doctor by her friends in the building where she lives. The CDSS provided different remission probabilities per treatment for each patient.

Participants arrived in groups of up to six, and were given an introductory session that covered the current state of depression treatment, the rationale for the development of an artificial intelligence-powered tool, current results of the artificial intelligence model and an introduction to the user interface of the tool. They were told that the standardised patients were playing patients who had used the tool to fill out questionnaires in the 'waiting room', but had limited knowledge of the tool.

Participants were paired with a research assistant, who guided them through a 10-min training session with the CDSS on a laptop. Participants then filled out a questionnaire recording their initial impressions of the tool. Each participant then interacted with all three standardised patients in a random order in three 10-min clinical scenarios. During scenarios participants were free to interact with a laptop computer running the CDSS. The laptop was angled at 45 degrees toward the participant, but could be freely moved to face the standardised patient. The CDSS had access to questionnaire results as well as the treatment algorithm with its integrated artificial intelligence tool. Participants were warned that as scenarios were only 10 min long, they should consider starting to use the CDSS roughly halfway through; however, they were also told that they had the freedom to use or ignore the CDSS as they saw fit. After each scenario, participants filled out a questionnaire about their experience using the CDSS.

After the scenarios, there was a 10-min structured interview with a research assistant in which participants were able to elaborate further on their experience. They were then asked to complete an anonymous 'exit' questionnaire summarising their experience using the tool and their opinion of its impact on the physician–patient interaction. The last step was a 10-question surprise quiz on the CANMAT 2016 Guidelines for Depression Treatment, intended to establish participant knowledge of guidelines. After each testing day, an unstructured debriefing session was held with all standardised patients. Although standardised patient feedback is often not standardised, standardised patients have been shown to effectively assess clinical skills,[17,19,20,34,36,37] which motivated us to consider standardised patient feedback when assessing the impact of the tool on the clinician–patient interaction. See Fig. 1 for a flowchart of tasks participants completed during the study.

Description of tool development and decision to use simulation centre testing

The development pathway of the Aifred system is that of a medical device. The first steps involved needs assessments, discussions with stakeholders (such as physicians and patients) and the creation of a prototype, which was reviewed by independent experts (six psychiatrists). Then, in a process mirroring that described in Trivedi et al,[31] programming of the prototype into a functional application was overseen by the clinical authors working on the project and tested by them, fake patient data was input into the system to test and refine it, and then data from real patients (in our case, data from patients in the studies used to train the machine-learning system) was used for testing and the development of simulation scenarios. Concurrently, as in Trivedi et al,[31] field testing with physicians (ongoing at present) has been used to collect feedback on the design and clinical validity and utility of a version of the tool without the artificial intelligence enabled (as the version of the tool with artificial intelligence enabled is a medical device that must only be used as part of clinical trials and related studies). The fact that our tool includes a novel artificial intelligence/ machine-learning component prompted further reflection on what studies were necessary to understand the impact of this novel component on the implementation of the CDSS. As a result, we decided we required a process evaluation, which, as discussed by Lamé and Dixon-Woods,[3] involves taking a 'look at how the intervention is implemented and received' and can be carried out, among other options, using a simulation setting. Simulation centres are beneficial not only for clinical tool assessment during development, but for simulation of realistic patient outcomes: a recent systematic review and meta-analysis of 33 studies found that simulation-based assessments involving healthcare professionals using technology-enhanced simulation in the context of patient care have been found to correlate positively with patient-related outcomes.[38] However, the quality of methods and reporting have been insufficient, a limitation we aimed to address by standardising our methods to previous research. Our development of a simulation centre study to conduct our process evaluation mirrors closely the method described by Colman et al[39] for developing simulation-based testing for healthcare spaces: as noted, we began with stakeholder engagement and needs assessment, and discussed the project and the simulations with a multidisciplinary team including computer scientists, clinicians, patients and people with research skills in fields such as anthropology. Clinical scenarios were then developed based on real patient data and situations that were likely to be encountered by the end-users of the CDSS. Standardised patients were then trained; an advantage of using the Steinberg Simulation Centre was that the standardised patients were professional actors skilled at preparing and standardising their performances, using a standard training process managed by simulation centre staff.[40] A testing day was then held as suggested by Colman et al,[39] with run-throughs of patient scenarios, a walk-through of the simulation space and a review of all training documents prepared for the testing day. The testing days were then held, with standardised patient and staff debriefings occurring each day, as suggested by Colman et al,[39] and this was then followed by data analysis and the creation of manuscripts for publication. We structured our analysis and reporting to assess some of the metrics of effective medical education as discussed by Dixon;[41] we chose medical education as a model given that the simulation centre experience did effectively act as a training session for use of the tool for physicians who participated. In this case, relevant areas of assessment as per Dixon,[41] were perception and opinion about the experience (often measured as
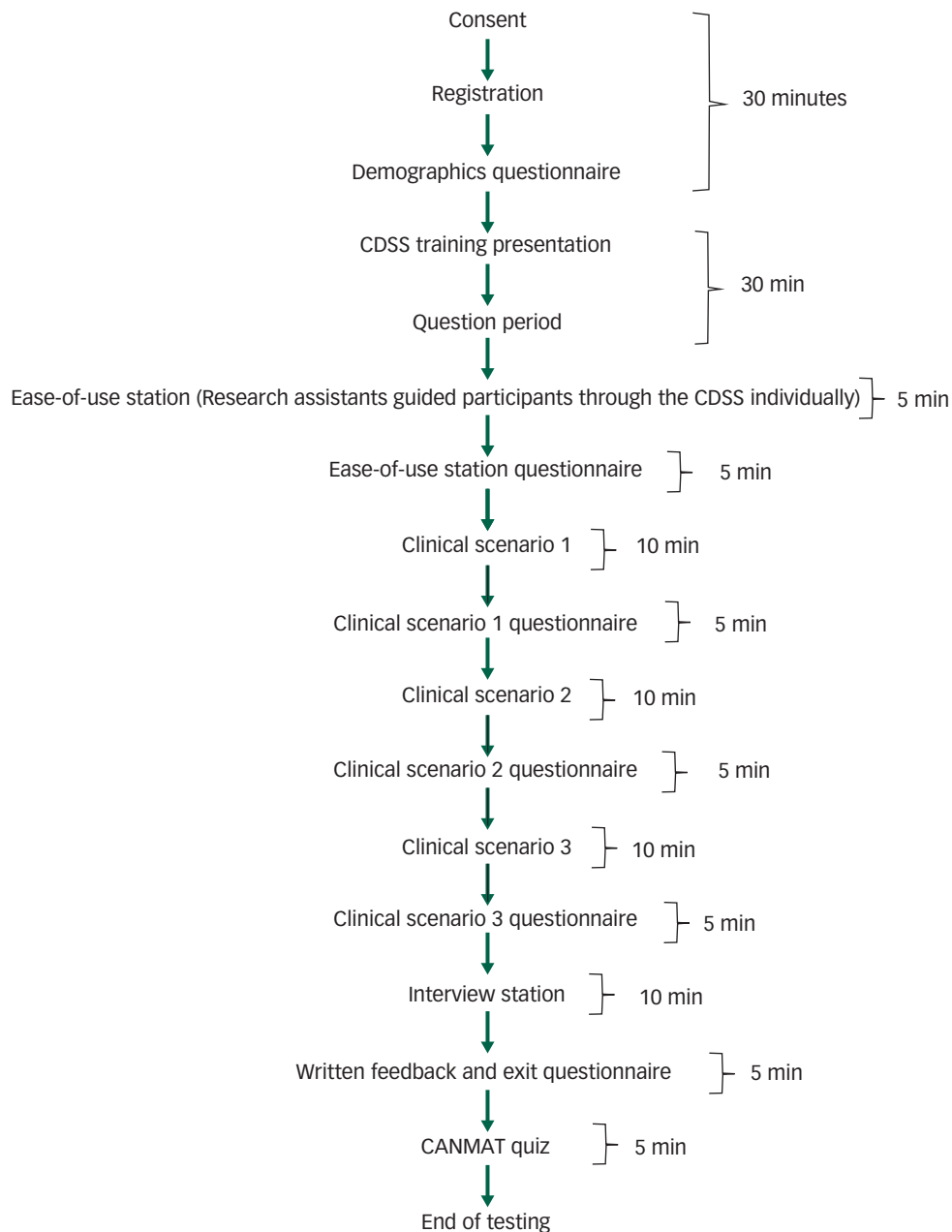
Fig. 1 Flowchart detailing the tasks participants completed during the study. CDSS, clinical decision support system.

satisfaction), knowledge or skills gained, and impact on clinical practice (with the latter only being inferred from responses clinicians gave about their likely future use of the tool).

Results

Results are derived from the registration and exit questionnaires, unless otherwise noted, and comment on participant satisfaction, knowledge and skills gained, and potential impact on clinical practice. Note that these are initial selected results meant to illustrate the utility of the simulation centre; full study results will be reported separately.

Twenty participants completed the study. Participants were nearly evenly split between psychiatry ($n = 11$) and family medicine ($n = 9$), with a wide age range (24–67 years, mean age 39.5 years) and practice experience (6 residents and the following breakdown

in experience for attending staff: 0–5 years: 4, 6–10 years: 2, 11–15 years: 2, 16–20 years: 4, ≥21 years: 2). The sample included participants practicing in hospital and community settings.

With respect to participant satisfaction and impact on the physician–patient interaction, 70% of participants felt that the artificial intelligence model assisted them in helping their patients better understand treatment (scoring ≥4 on a scale of 1–5, with higher values representing greater confidence). Sixty-five per cent of participants felt it helped improve patient trust in the treatment (scoring ≥4 on a scale of 1–5). Fifty per cent of participants felt that the application provided them with richer information to discuss with their patients (scoring ≥4 on a scale of 1–5). Forty-five per cent of participants reported that using the application made the interaction with patients feel less personal or that it interfered with their interview (scoring ≥4 on a scale of 1–5). Seventy per cent of participants felt the remission probabilities provided by the model were reasonable overall.

In terms of potential impact on clinical practice, 50% of participants thought they would use the CDSS for all of their patients with MDD, with an additional 40% (therefore 90% overall) stating they would use it for more complex or treatment-resistant patients. Sixty per cent of participants trusted that the artificial intelligence could help them choose treatments (scoring ≥4 on a scale of 1–5). Eighty per cent of participants felt that the information on the treatment selection page in the application (which indicated CANMAT-recommended treatments, their usual doses and the artificial intelligence predictions) contained information that was clinically useful (scoring ≥4 on a scale of 1–5). This suggests that the information contained in the tool could augment clinician knowledge during their interactions with patients. See Table 1 for a summary of results.

Before the simulation, 75% of participants reported that they would realistically use the application in clinic for 5 min or less during a session. Forty per cent of participants reported that the application would save them time (scoring ≥4 on a scale of 1–5), and 30% felt the application would neither save nor cost them time (scoring 3 on a scale of 1–5), indicating potential feasibility in a real, busy clinical environment. This was corroborated by the fact that, in the majority of scenarios, the participants were able to successfully navigate through the application within the short time provided. In a questionnaire administered right after each clinical scenario, 61.7% of participants reported that using the application 'took some adjustment, but […] worked well'. Standardised patients provided valuable feedback, such as noting that some participants turned the computer screen toward them during the session, 'inviting them in' to engage with the tool. This seemed to be linked to acceptability of the tool's presence on the part of the standardised patients. They also commented on the importance of the clinician's manner and rapport building skills, such as warmth and ability to engage them in their care.

Discussion

We will now reflect on the use of simulation for testing the effect of new technologies on the physician–patient interaction. Our initial results demonstrate that a majority of clinicians were satisfied with the use of the CDSS. At the end of the simulation, most clinicians could see themselves using the tool for at least a subset of their patients with depression, suggesting the feasibility of using the tool to achieve real-world impact. No major threats to the quality of the physician–patient interaction were identified, and we illustrated several ways in which the tool might enhance the interaction, as well as tools clinicians can use to better integrate the CDSS into a session.

Our sample of 20 participants was diverse with respect to career stage and practice environment, which increases confidence in the generalisability of our results. The sample size reflects recruitment feasibility. The largest barrier to recruitment was clinical duties and, for residents, concerns about not being released to participate. Being able to offer more testing days, as well as departmental approval for residents' participation, may have increased recruitment. A challenge with simulation is that running participants in groups on predefined days is necessary given the need to ensure room and standardised patient availability.

Using simulation-based testing allowed us to observe interactions that would not have been easily accessible in other settings. As noted, some participants tended to turn the laptop toward their standardised patient. Standardised patients referred to this as participants 'inviting them in'; this behaviour seemed to be important in determining their experience of the tool. Standardised patient feedback and our observations of sessions revealed that traditional aspects of the physician–patient interaction, such as clinician warmth, body language and ability to engage the patient, were also important in determining

the standardised patient experience, suggesting that the impact of a new technology may depend on clinicians' baseline ability to build rapport with their patients. This merits further investigation in a clinical environment. Self-report from clinicians also revealed important effects of the CDSS on the physician–patient interaction, such as the perceived utility of the tool in helping them better explain and increase trust in treatment. This interplay of observations of clinician behaviour, clinician self-report and standardised patient experiences provided fundamentally different information than would have been obtained through clinician self-report alone. These observations will influence clinician training provided in future clinical studies, resulting in more focus on how clinicians can engage the patient with the tool in-session and use it to provide more information and enhance patient trust.

External validity is a concern when using simulation-based testing.[3] For example, several participants noted in written comments and during interviews that the 10-min training session was insufficient and that they would likely have become more comfortable with the CDSS with more time. However, external validity may depend on research aims.[3] In our case, the aim was to see if the application was intuitive to use with minimal training, and, as noted, the majority of participants felt the tool took some adjustment but worked well. Similarly, the 10-min clinical session length was felt by multiple participants to be too short. We initially hypothesised that most clinicians would want a tool that they could use in 5 min, and this was supported by the finding that, at baseline, 75% of participants could see themselves using the tool for 5 min or less. Having short sessions, in which most participants used the tool in the latter half of the session, allowed us to determine that it is possible to use the tool in a meaningful way within this time constraint. As such, our research aims were well suited to simulation work.

The use of the simulation environment – and crucially, of standardised patients – to test the impact of technology on the physician–patient interaction is both practically useful and important as it allows direct observation of clinician interaction with a new tool before patient studies. This method provides multiple points of observation, allowing for an informative and multifaceted dataset that can inform the development of tools and training materials. Evaluating the ease with which new technology is used and integrated into clinical practice is a key step in the proper development and implementation of novel clinical tools, and is a useful prelude to more longitudinal studies on the impact of these tools on the clinician–patient relationship.

With respect to engaging patients in shared decision-making in the context of CDSS use, during this study physicians could have chosen any number of approaches. For example, they could have started by turning the screen toward the patient; kept the screen toward themselves while discussing the treatments and artificial intelligence results with the patient; or referred to the CDSS with the screen turned toward them, and then put it away and discussed treatment with the patient without explicitly discussing the CDSS and its results. The finding that standardised patients were more accepting of the tool when clinicians turned the screen toward them and 'invited them in' is not surprising in and of itself. However, it is instructive as it provides a concrete and simple behaviour that seems to have a significant impact on patient experience, the promotion of which can be included as part of the training for clinicians using the tool in the clinic or as part of coming clinical studies. It is also a finding that helps us determine which of the possible clinician behaviours in response to the tool would be most likely to be supportive of patients feeling engaged in decision-making. In addition, having actually observed the importance of this behaviour under simulation conditions may potentially help convince clinicians to adopt it.

In previous research, Trivedi et al[42] identified several barriers to implementation of a computerised decision support system. These

The left margin has vertical text with DOI and publisher info, right margin has article title.

Table 1	Study results by category				
Category	Question		Scale	Percentages	Summary
Participant satisfaction	The probabilities produced by the model, overall, were:		Too optimistic	15%	70% of participants felt remission probabilities were reasonable.
			Reasonable	70%	
			Too pessimistic	15%	
	What impact do you think the predictive model, in particular, had on the patient–clinician interaction? Please rate your agreement.	I felt I could use the model to help my patient better understand treatment:	Strongly agree	15%	70% of participants felt that the artificial intelligence model helped them to help their patients better understand treatment.
			Somewhat agree	55%	
			Unsure	10%	
			Somewhat disagree	20%	
			Strongly disagree	0%	
		The numbers provided by the model improved trust in the treatment:	Strongly agree	15%	65% of participants felt the numbers provided by the model improved trust in the treatment.
			Somewhat agree	50%	
			Unsure	20%	
			Somewhat disagree	15%	
			Strongly disagree	0%	
		The model provided us with more rich information to discuss:	Strongly agree	10%	50% of participants felt the model provided them with richer information to discuss with patients.
			Somewhat agree	40%	
			Unsure	30%	
			Somewhat disagree	15%	
			Strongly disagree	5%	
		The application made the interaction less personal:	Strongly agree	20%	45% of participants felt the application made interaction with patients less personal.
			Somewhat agree	25%	
			Unsure	10%	
			Somewhat disagree	35%	
			Strongly disagree	10%	
		The application interfered with my patient interview:	Strongly agree	20%	45% of participants felt the application interfered with their patient interview.
			Somewhat agree	25%	
			Unsure	10%	
			Somewhat disagree	40%	
			Strongly disagree	5%	
Knowledge and skills gained		Based on your overall experience today, how much do you trust the predictive model to help you choose treatments for depression (1 being 'very little' and 5 being 'very much')?	5	10%	60% of participants trusted the predictive model to help choose treatments.
			4	50%	
			3	20%	
			2	15%	
			1	5%	
		Rate your agreement with the following statement: The information on the page where I had to select treatment was clinically useful:	Strongly agree	25%	80% of participants felt the information on the treatment selection page was clinically useful.
			Somewhat agree	55%	
			Unsure	10%	
			Somewhat disagree	10%	
			Strongly disagree	0%	

(*Continued*)

Table 1 (Continued)

Category	Question	Scale	Percentages	Summary
Potential impact on clinical practice	Based on your experience today, do you think using the application would cost you significant time (1 being 'cost you significant time' and 5 being 'save you significant time'):	5	5%	40% of participants felt the application would save them time, and 30% felt the application would neither cost nor save time.
		4	35%	
		3	30%	
		2	25%	
		1	5%	
	You would use the application	For all patients with depression	50%	50% of participants thought they would use the application for all patients with depression, and an additional 40% thought they would use the application for more complex or treatment-resistant patients. Thus, 90% of participants said they would use the application with at least some of their depression patients.
		Only for the most severe patients	5%	
		Only for patients where one treatment has failed	25%	
		Only for patients where more than one treatment has failed	5%	
		Not at all	0%	
		Other:		
		Patients with confounding factors	5%	
		To review patient info	5%	
		Likely not	5%	

included concerns about the time required to use the system in practice, technical challenges relating to computer literacy, and the need for physicians to be involved in the development of the tool and to have the ability to override system recommendations. Accordingly, we designed the tool to ensure physician autonomy by allowing physicians to select any treatment or action they deemed appropriate. We included physicians in the design and iterative ongoing design process, of which this simulation centre study is a part. And finally, we designed the tool to be easy to use quickly and intuitively during a patient encounter. As noted, physicians were able, after a short training session, to use the tool effectively within a short clinical encounter. This study aimed to assess, as part of the development of this tool, if we are on the right track in addressing some of the barriers previously noted in CDSS implementation; the present results provide preliminary indications that this is the case, which will be further assessed in a clinical feasibility study.

This study has a number of limitations and serves as only an initial step in the examination of the effect of this tool on the clinical process, with its main purpose being the identification of significant problems in the patient–clinician encounter when using the tool as well as the refinement of training materials for further clinical studies. In line with Dixon's[41] comments when discussing the validity of medical education evaluation, one cannot assume that changes in physician knowledge or skills, or satisfaction with the training or the tool itself, will directly lead to improved patient outcomes; furthermore, one cannot assume that the impressions physicians had of the tool with respect to its potential effect on their practice would be borne out once they begin using it in clinic. The present study does, however, help establish that physicians seem open to trying this tool in clinic, that they can be easily trained to use it in a manner they find satisfactory and that there is some agreement among physicians that the tool has potential clinical utility. As such, the next step will be to conduct a feasibility study of the tool in a longitudinal manner in clinic, followed by a randomised control study aimed at assessing tool effectiveness and safety. The largest drawback of this simulation centre study, with respect to assessment of the effect of the tool on the patient–clinician interaction, is that it is impossible, in this setting, to assess longitudinal effects on the patient–clinician relationship, hence raising the importance of conducting a longitudinal feasibility study in clinic before large-scale clinical trials.

Although this study does not evaluate the effectiveness of this tool, it has provided valuable insights into how clinicians may use this type of tool and how the tool, and the training provided to clinicians who use it, may be further developed to increase the chance that it will have a positive impact on patient care.

David Benrimoh [ORCID], Department of Psychiatry, McGill University, Canada; Aifred Heath Inc., Montreal, Canada; and Faculty of Medicine, McGill University, Canada;
Myriam Tanguay-Sela [ORCID], Montreal Neurological Institute, McGill University, Canada; and Aifred Health Inc., Montreal, Canada; **Kelly Perlman** [ORCID], Douglas Mental Health University Institute, Montreal, Canada; and Aifred Health Inc., Montreal, Canada;
Sonia Israel, Aifred Health Inc., Montreal, Canada; **Joseph Mehltretter**, Department of Computer Science, University of Southern California, Los Angeles, USA; and Aifred Health Inc., Montreal, Canada; **Caitrin Armstrong**, School of Computer Science, McGill University, Canada; and Aifred Health Inc., Montreal, Canada; **Robert Fratila**, Aifred Health Inc., Montreal, Canada; **Sagar V. Parikh**, Department of Psychiatry, University of Michigan, USA; **Jordan F. Karp**, Department of Psychiatry, University of Pittsburgh, USA; **Katherine Heller**, Department of Statistical Science, Duke University, USA; **Ipsit V. Vahia**, Department of Psychiatry, McLean Hospital/Harvard University, USA; **Daniel M. Blumberger** [ORCID], School of Psychiatry, University of Toronto, Canada; **Sherif Karama**, Department of Psychiatry, McGill University, Canada; **Simone N. Vigod** [ORCID], Department of Psychiatry, University of Toronto, Canada; **Gail Myhr**, Department of Psychiatry, McGill University, Canada; **Ruben Martins**, Douglas Mental Health University Institute, Montreal, Canada; and Department of Psychiatry, McGill University, Canada; **Colleen Rollins** [ORCID], Department of Psychiatry, University of Cambridge, UK; and Aifred Health Inc., Montreal, Canada; **Christina Popescu** [ORCID], Douglas Mental Health University Institute, Montreal, Canada; and Aifred Health Inc., Montreal, Canada; **Eryn Lundrigan** [ORCID], Department of Anatomy and Cell Biology, McGill University, Canada; **Emily Snook**, Faculty of Medicine, University of Toronto, Canada; **Marina Wakid**, Douglas Mental Health University Institute, Montreal, Canada; **Jérôme Williams** [ORCID], Faculty of Medicine, McGill University, Canada;

Ghassen Soufi, Faculty of Medicine, McGill University, Canada; **Tamara Perez** (iD), Department of Experimental Medicine, McGill University, Canada; **Jingla-Fri Tunteng**, School of Physical and Occupational Therapy, McGill University, Canada; **Katherine Rosenfeld**, Department of Psychology, McGill University, Canada; **Marc Miresco**, Department of Psychiatry, McGill University, Canada; **Gustavo Turecki**, Douglas Mental Health University Institute, Montreal, Canada; and Department of Psychiatry, McGill University, Canada; **Liliana Gomez Cardona**, Douglas Mental Health University Institute, Montreal, Canada; and Department of Psychiatry, McGill University, Canada; **Outi Linnaranta**, Douglas Mental Health University Institute, Montreal, Canada; and Department of Psychiatry, McGill University, Canada; **Howard C. Margolese**, Department of Psychiatry, McGill University, Canada

Correspondence: David Benrimoh. Email: david.benrimoh@mail.mcgill.ca

First received 23 Mar 2020, accepted 29 Sep 2020

Acknowledgements

Funding

Author contributions

D.B. worked on conceptualizing and running the study and writing the protocol, oversaw analysis, and worked on creating and revising the draft manuscript. M.T.-S. worked on running the study, organizing research assistants, oversaw analysis, and worked on creating and revising the draft manuscript. K.P. and S.I. helped conceptualize the study and write the protocol, collected data and helped revise the manuscript. K.P. also contributed to data analysis. J.M., C.A. and R.F. created, in collaboration with D.B., the AI model tested in the study and designed how it would report information. They also helped revise the manuscript. S.V.P., J.F.K., K.H. and I.V.V. provided comments on study protocol and measures, provided guidance on the analysis, and helped significantly revise the manuscript. In addition, S.V.P. originated the idea for this format of manuscript and provided a number of references. J.F.K. provided part of the data used to train the AI model. S.K., S.N.V., G.M., R.M. and D.M.B. helped assess the clinical validity of the treatment algorithm into which the AI was inserted and helped revise the manuscript. C.P., E.L., M.W., J.W., G.S., T.P., J.-F.T. and K.R. were all research assistants who assisted in data collection and analysis, and revised the manuscript. C.R. was a research assistant who helped write the original protocol, assisted with data analysis and revised the manuscript. E.S. was a research assistant who assisted with data analysis and revising the manuscript. M.M. and G.T. assisted in the development of the original protocol, provided research questions for data analysis and also helped revise the manuscript. L.G.C. and O.L. provided comments on the data analysis and helped revise the manuscript. H.C.M. helped conceptualize the study and produce the original protocol, and oversaw data analysis. He also significantly revised the manuscript.

Declaration of interest

References

1   Rosenfeld A, Benrimoh D, Armstrong C, Mirchi N, Langlois-Therrien T, Rollins C, et al. Big data analytics and AI in mental healthcare. ArXiv [Preprint]. 2019. Available from: https://arxiv.org/abs/1903.12071.

2   Nolan P, Badger F. Aspects of the relationship between doctors and depressed patients that enhance satisfaction with primary care. *J. Psychiatr Ment Health Nurs* 2005; **12**(2): 146–53.

3   Lamé G, Dixon-Woods M. Using clinical simulation to study how to improve quality and safety in healthcare. *BMJ Simul Technol Enhanc Learn* 2020; **6**: 87–94.

4   Patton MQ. Enhancing the quality and credibility of qualitative analysis. *Health Serv Res* 1999; **34**(5): 1189–208.

5   Bromet E, Andrade LH, Hwang I, Sampson NA, Alonso J, De Girolamo G, et al. Cross-national epidemiology of DSM-IV major depressive episode. *BMC Med* 2011; **9**(1): 90.

6   Ferrari AJ, Charlson FJ, Norman RE, Patten SB, Freedman G, Murray CJ, et al. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLoS Med* 2013; **10**(11): e1001547.

7. World Health Organization. *Depression and Other Common Mental Disorders: Global Health Estimates*. World Health Organization, 2017 (http://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSDMER-2017.2-eng.pdf?sequence=1).

8   Williams SZ, Chung GS, Muennig PA. Undiagnosed depression: a community diagnosis. *SSM Popul Health* 2017; **3**: 633–8.

9   Warden D, Rush AJ, Trivedi MH, Fava M, Wisniewski SR. The STAR* D project results: a comprehensive review of findings. *Curr Psychiatry Rep* 2007; **9**(6): 449–59.

10 Benrimoh D, Fratila R, Israel S, Perlman K, Mirchi N, Desai S, et al. Aifred health, a deep learning powered clinical decision support system for mental health. In *The NIPS'17 Competition: Building Intelligent Systems*: 251–87. Springer, 2018.

11 Greden JF, Parikh SV, Rothschild AJ, Thase ME, Dunlop BW, DeBattista C, et al. Impact of pharmacogenomics on clinical outcomes in major depressive disorder in the GUIDED trial: a large, patient-and rater-blinded, randomized, controlled study. *J Psychiatr Res* 2019; **111**: 59–67.

12 Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry* 2016; **3**(3): 243–50.

13 Webb CA, Trivedi MH, Cohen ZD, Dillon DG, Fournier JC, Goer F, et al. Personalized prediction of antidepressant v. placebo response: evidence from the EMBARC study. *Psychol Med* 2019; **49**(7): 1118–27.

14 Kennedy SH, Lam RW, McIntyre RS, Tourjman SV, Bhat V, Blier P, et al. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 clinical guidelines for the management of adults with major depressive disorder: section 3. Pharmacological treatments. *Can J Psychiatry* 2016; **61**(9): 540–60.

15 Mehltretter J, Rollins C, Benrimoh D, Fratila R, Perlman K, Israel S, et al. Analysis of features selected by a deep learning model for differential treatment selection in depression. *Front Artif Intell* [Epub ahead of print] 21 Jan 2020. Available from: https://doi.org/10.3389/frai.2019.00031.

16 Kroenke KL, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001; **16**(9): 606–13.

17 Kessler RC, Magee WJ. Childhood adversities and adult depression: basic patterns of association in a US national survey. *Psychol Med* 1993; **23**(3): 679–90.

18 Lam RW, Mcintosh D, Wang J, Enns MW, Kolivakis T, Michalak EE, et al. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 clinical guidelines for the management of adults with major depressive disorder: section 1. Disease burden and principles of care. *Can J Psychiatry* 2016; **61**(9): 510–23.

19 World Health Organization. *Depression and Other Common Mental Disorders*. World Health Organization, 2017 (https://www.who.int/mental_health/management/depression/prevalence_global_health_estimates/en/).

20 Cipriani A, Furukawa TA, Salanti G, Chaimani A, Atkinson LZ, Ogawa Y, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Focus* 2018; **16**(4): 420–9.

21 Iniesta R, Hodgson K, Stahl D, Malki K, Maier W, Rietschel M, et al. Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. *Sci Rep* 2018; **8**: 5530.

22 Rush AJ, Trivedi MH, Stewart JW, Nierenberg AA, Fava M, Kurian BT, et al. Combining medications to enhance depression outcomes (CO-MED): acute and long-term outcomes of a single-blind randomized study. *Am J Psychiatry* 2011; **168**(7): 689–701.

23 Trivedi MH, McGrath PJ, Fava M, Parsey RV, Kurian BT, Phillips ML, et al. Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): rationale and design. *J Psychiatr Res* 2016; **78**: 11–23.

24 Trivedi MH, Kocsi JH, Thase ME, Morris DW, Wisniewski SR, Leon AC, et al. REVAMP - research evaluating the value of augmenting medication with psychotherapy: rationale and design. *Psychopharmacol Bull* 2008; **41**(4): 5–33.

25 Lenze EJ, Mulsant BH, Blumberger DM, Karp JF, Newcomer JW, Anderson SJ, et al. Efficacy, safety, and tolerability of augmentation pharmacotherapy with aripiprazole for treatment-resistant depression in late life: a randomised, double-blind, placebo-controlled trial. *Lancet* 2015; **386**(10011): 2404–12.

26 Hopwood M. The shared decision-making process in the pharmacological management of depression. *Patient* 2020; **13**: 23–30.

27 Trivedi MH, Rush AJ, Crismon ML, Kashner TM, Toprac MG, Carmody TJ, et al. Clinical results for patients with major depressive disorder in the Texas Medication Algorithm Project. *Arch Gen Psychiatry* 2004; **61**(7): 669–80.

28 Adli M, Wiethoff K, Baghai TC, Fisher R, Seemüller F, Laakmann G, 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* 2017; **20**(9): 721–30.

29 Dobscha SK, Corson K, Hickam DH, Perrin NA, Kraemer DF, Gerrity MS. Depression decision support in primary care: a cluster randomized trial. *Ann Intern Med* 2006; **145**(7): 477–87.

30 Rollman BL, Hanusa BH, Lowe HJ, Gilbert T, Kapoor WN, Schulberg HC. A randomized trial using computerized decision support to improve treatment of major depression in primary care. *J Gen Intern Med* 2002; **17**(7): 493–503.

31 Trivedi MH, Kern JK, Grannemann BD, Altshuler KZ, Sunderajan P. A computerized clinical decision support system as a means of implementing depression guidelines. *Psychiatr Serv* 2004; **55**(8): 879–85.

32 Kurian BT, Trivedi MH, Grannemann BD, Claassen CA, Daly EJ, Sunderajan P. A computerized decision support system for depression in primary care. *Prim Care Companion J Clin Psychiatry* 2009; **11**(4): 140.

33 Harrison P, Carr E, Goldsmith K, Young AH, Ashworth M, Fennema D, et al. 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* 2020; **10**(5): e035905.

34 Beullens J, Rethans JJ, Goedhuys J, Buntinx F. The use of standardized patients in research in general practice. *Fam Pract* 1997; **14**(1): 58–62.

35 Shirazi M, Sadeghi M, Emami A, Kashani AS, Parikh S, Alaeddini F, et al. Training and validation of standardized patients for unannounced assessment of physicians' management of depression. *Acad Psychiatry* 2011; **35**(6): 382–7.

36 Bokken L, Linssen T, Scherpbier A, Vleuten CVD, Rethans JJ. Feedback by simulated patients in undergraduate medical education: a systematic review of the literature. *Med Educ* 2009; **43**(3): 202–10.

37 Park JH, Son JY, Kim S, May W. Effect of feedback from standardized patients on medical students' performance and perceptions of the neurological examination. *Med Teach* 2011; **33**(12): 1005–10.

38 Brydges R, Hatala R, Zendejas B, Erwin PJ, Cook DA. Linking simulation-based educational assessments and patient-related outcomes: a systematic review and meta-analysis. *Acad Med* 2015; **90**(2): 246–56.

39 Colman N, Doughty C, Arnold J, Stone K, Reid J, Dalpiaz A, et al. Simulation-based clinical systems testing for healthcare spaces: from intake through implementation. *Adv Simul* 2019; **4**(1): 19.

40 Mueller CL, Cyr G, Bank I, Bhanji F, Birnbaum L, Boillat M, et al. The Steinberg Centre for Simulation and Interactive Learning at McGill University. *J Surg Educ* 2017; **74**(6): 1135–41.

41 Dixon J. Evaluation criteria in studies of continuing education in the health professions: a critical review and a suggested strategy. *Eval Health Prof* 1978; **1**(2): 47–65.

42 Trivedi MH, Daly EJ, Kern JK, Grannemann BD, Sunderajan P, Claassen CA. Barriers to implementation of a computerized decision support system for depression: an observational report on lessons learned in "real world" clinical settings. *BMC Med Inform Decis Making* 2009; **9**: 6.