Integrating AI Recommendations into the Pharmacologic Management of Major Depressive Disorder

Maia Jacobs
Harvard University
Cambridge, MA 02138, USA
mjacobs@g.harvard.edu

Roy H. Perlis
Massachusetts General Hospital
Cambridge, MA 02114, USA
rperlis@mgh.harvard.edu

Melanie F. Pradier
Harvard University
Cambridge, MA 02138, USA
melanie@seas.harvard.edu

Finale Doshi-Velez
Harvard University
Cambridge, MA 02138, USA
finale@seas.harvard.edu

Elizabeth Mynatt
Georgia Institute of Technology
Atlanta, GA 30308
mynatt@gatech.edu

Krzysztof Z. Gajos
Harvard University
Cambridge, MA 02138, USA
kgajos@eecs.harvard.edu

CSCW'19,, Nov 9–13, 2019, Austin, TX, USA
ACM 978-1-4503-6819-3/20/04.

Abstract

AI predictions provide an important opportunity to support clinicians during complex decision-making processes. One such process is selecting treatments for major depressive disorder (MDD). Towards the goal of implementing AI models that make MDD treatment recommendations, we have designed a factorial vignette study to assess how recommendations and explanations may influence clinician's treatment decisions. We report on our initial data analysis, evaluating the influence of incorrect predictions on antidepressant selection. We found that recommendation correctness had a significant effect on treatment selection accuracy.

Author Keywords

Clinical decision support; major depressive disorder.

Introduction

AI-based decision-support tools are expected to transform many aspects of healthcare, helping to diagnose illnesses and determine longitudinal health risks [5, 3]. Predictive models may also be able to assist clinicians in making complex treatment decisions, but to date few tools have been designed or implemented in clinical settings to facilitate a collaborative decision-making process between clinicians and AI systems [11].

One context that may benefit from the implementation of predictive models is the treatment selection process for major depressive disorder (MDD). The pharmacologic management of MDD currently involves trial and error. Currently, 2/3 of patients diagnosed with MDD fail to reach remission with their initial treatment, and 1/3 of the patient population do not remit despite up to four antidepressant trials [10]. The treatment decision process is also complicated due to the limited guidance available for clinicians regarding antidepressant medications, especially when selecting a secondary or tertiary treatment for a patient.

Researchers are working to develop predictive models to support antidepressant medication selection [4, 9]. These models are trained on electronic medical record data to predict treatment success for an individual diagnosed with MDD. However, questions remain about how such models may be implemented in clinical practice so that they are both usable and useful to clinicians.

Towards the goal of supporting the psychiatric decision making for the treatment of MDD, this user study looks at how different representations of the model's output may influence clinical decisions. In the broader study, we are examining the influence of multiple factors on decision-making, including recommendation accuracy, explanation styles, and clinicians' familiarity with the recommended treatments. Here, we present results from the study demonstrating the influence of inaccurate recommendations on treatment decisions.

Methods

The focus of this study was to explore how AI recommendations may influence MDD treatment selections, and the possible consequences of inaccurate recommendations. To study the effects of recommendation correctness on the dependent variables (accuracy, confidence, and perceived helpfulness), we used a factorial vignette survey design to study clinicians' decisions in a series of hypothetical patient scenarios with systematically varied AI recommendations [1]. The survey includes the following variables:

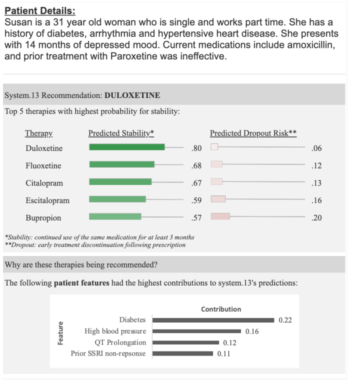
Random Variable

1. **Patient scenario**: We first worked with two psychiatrists to develop a series of realistic patient situations. Scenarios were randomly displayed. For every vignette (defined as a set of the independent variable conditions), a participant could see any of the five scenarios.

Independent Variables

1. **Recommendation correctness:** No recommendation, correct, incorrect. Correct and incorrect recommendations for each patient scenario were determined by experts in psychopharmacology. While we highlighted a single recommendation in each vignette, we also showed a top-5 list of recommended treatment options, as there are often several reasonable treatment options for a person diagnosed with MDD. For incorrect recommendation conditions, only the top recommendation was incorrect.

2. **Explanation types:** None, placebo, rule-based, feature-importance. With no explanation, a participant only sees the treatment recommendations. Placebo explanations state only that "recommendations are based on patients' ICD-9 codes". We included placebo explanations to distinguish between effects caused by the visibility of a explanation and the content of a explanation. Finally, we included rule-based and feature-based explanations, as both of these styles have been successfully implemented in



Patient Details:
Susan is a 31 year old woman who is single and works part time. She has a history of diabetes, arrhythmia and hypertensive heart disease. She presents with 14 months of depressed mood. Current medications include amoxicillin, and prior treatment with Paroxetine was ineffective.

System.13 Recommendation: **DULOXETINE**

Top 5 therapies with highest probability for stability:

Therapy	Predicted Stability*	Predicted Dropout Risk**
Duloxetine	.80	.06
Fluoxetine	.68	.12
Citalopram	.67	.13
Escitalopram	.59	.16
Bupropion	.57	.20

**Stability: continued use of the same medication for at least 3 months*
***Dropout: early treatment discontinuation following prescription*

Why are these therapies being recommended?

The following **patient features** had the highest contributions to system.13's predictions:

Figure 1: Sample vignette with a patient scenario, AI recommendations, and feature-based explanation.

other contexts [2], allowing us to examine if design of explanations in non-medical domains may be useful in the design of clinical systems.

3. **Treatment types:** Common, less common. Another question we are exploring in this survey is how the use of AI predictions change if the top recommendation is a more commonly prescribed treatment (selective serotonin reuptake inhibitors), or less commonly prescribed [7].

Dependent Variables

1. **Treatment selection accuracy:** To determine accuracy scores, we worked with five experts in psychopharmacology to rate 24 antidepressant treatment options for each patient scenario. They used a 4-point rating scale: 0=worst choice, 1=poor choice, 2=reasonable choice,3=best choice. We used the mode of their ratings to assign an accuracy score for each treatment in each patient scenario.

2. **Treatment selection confidence:** In each vignette, after selecting a treatment, participants were asked "How confident are you with this decision" using a 5-point Likert scale (1=not at all confident, 5=extremely confident).

3. **Perceived helpfulness of the AI system:** For each vignette, participants were asked to rate how helpful the AI system was in making their decision, using a 5-point Likert scale (1=Not at all, 5=A great deal).

The survey was deployed using Qualtrics and used a within subject design, so that participants saw every combination of the three independent variables. Each participant saw the conditions in a different random order. The data was analyzed using JMP Pro v14. Figure 1 shows an example interface with a treatment recommendation and explanation.

Participants

Using social media and snowball sampling, we received 240 survey responses. 20 were removed due to ineligibility and 220 were included in the analysis. Reasons for ineligibility included having <1 year of experience prescribing antidepressant treatments or not providing a medical specialty. We also decided to remove responses from outside of the United States due to the small response rate and possible differences in training and treatment selection processes.

Of the 220 participants, medical specialties included psychiatry (n=195), primary care (n=18), and other medical specialties (n=7). Participant ages ranged from 27–81 (mean=42.5). Participants years of experience prescribing antidepressant treatments ranged from 1–50 years (mean=12.1).

Preliminary Findings

In this analysis, we look at the influence of correct and incorrect recommendations (across all explanation styles) on three variables: treatment decision accuracy, confidence in the decision, and perceived helpfulness of the recommendation. These results are illustrated in Figure 2.

Our results indicate that recommendation correctness had a significant main effect on treatment selection accuracy (F(2,2820)=14.736, p<.0001). Tukey post hoc tests suggest that accuracy scores when incorrect recommendations were given (mean=1.573; sd=.744) were significantly lower than accuracy scores with no recommendation (mean=1.691; sd=.809; p=.002) and with correct recommendations (mean=1.754; sd=.808; p<.0001). Accuracy scores with correct recommendations were higher than accuracy score with no rec-

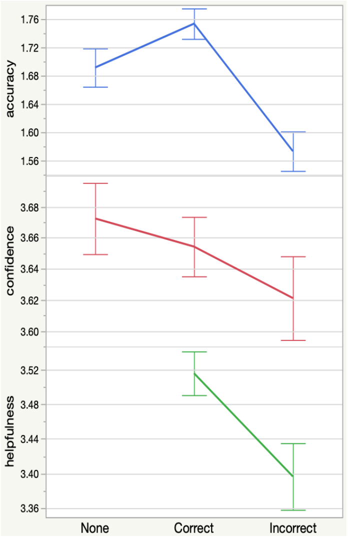


Figure 2: From top to bottom: the accuracy scores, confidence scores, and perceived helpfulness scores for vignettes with no recommendation (left), correct recommendations, and incorrect recommendations (right).

ommendations, but this difference was not statistically significant (p=.138).

We next look at participants' subjective responses (confidence and perceived helpfulness). Differences in perceived helpfulness of correct recommendations (mean=3.515; sd=1.052) and incorrect recommendations (mean=3.397; sd=1.133) were statistically significant (F(1,2410)=12.448, p=.0004). Differences in participants' confidence when provided no recommendations (mean=3.672; sd=.759), correct recommendations (mean=3.654; sd=.802) and incorrect recommendations (mean=3.621; sd=.799) were not statistically significant (F(2,3509)=1.653, p=.192).

Implications

An important consideration when implementing AI predictions into clinical practice is the potential influence of inaccurate information on the decision process. Using MDD as a case study, we compare how correct and incorrect recommendations change treatment decision accuracy, confidence in the decision, and perceived helpfulness of the recommendation.

We found that participants did rate incorrect recommendations as less helpful, a subjective indication that participants noticed a difference between the recommendation types. While participants' confidence in their treatment selection did decline slightly with incorrect recommendations, this effect was not significant. Looking at these results together, we see that while participants saw incorrect recommendations as less helpful, and remained confident in their decisions, performance ultimately declined. These results help to quantitatively show the potential implications of incorrect recommendations on clinical decisions.

Our results suggest that models need to be transparent in their limitations, highlighting situations in which the AI prediction may not be accurate or valid. However, even recommendations considered less helpful may influence treatment decisions. Therefore, displaying the limitations of an AI system may not be enough to optimize performance in human-AI collaboration. These findings are consistent with emerging results from other researchers (e.g., [8]). We believe there is a promising research agenda considering the role of design friction when creating AI interfaces and interactions. Design frictions, or points of difficulty in an interaction design, can improve understanding of a technology and encourage a person to reflect on their behavior or decision [6]. Design frictions may be a useful framework for moving away from the development of glanceable displays, which may lead to over-trust in the model, and support critical thinking with, and about, an AI prediction. Of course, inherent tensions exist, as design frictions purposefully slow down the decision making process, and medical decisions often take place in a time-critical environment. Open questions remain about how to best design for human-AI collaboration in a way that promotes critical thinking and reflection while remaining usable in a time-critical context.

Acknowledgements

We would like to thank all study participants. This work was supported by the Harvard Data Science Initiative Postdoctoral Fellow Research Fund.

REFERENCES

1. Lawrence H. Ganong and Marilyn Coleman. 2006. Multiple segment factorial vignette designs. *Journal of Marriage and Family* 68, 2 (2006), 455–468. DOI: http://dx.doi.org/10.1111/j.1741-3737.2006.00264.x
2. Riccardo Guidotti, Anna Monreale, and Salvatore Ruggieri. 2018. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (2018).

3. Longfei Han, Senlin Luo, Jianmin Yu, Limin Pan, and Songjing Chen. 2015. Rule extraction from support vector machines using ensemble learning approach: An application for diagnosis of diabetes. *IEEE Journal of Biomedical and Health Informatics* 19, 2 (2015), 728–734. DOI: http://dx.doi.org/10.1109/JBHI.2014.2325615

4. Michael C Hughes, Gabriel Hope, Leah Weiner, Thomas H Mccoy, Roy H Perlis, Erik Sudderth, and Finale Doshi-velez. 2018. Semi-Supervised Prediction-Constrained Topic Models. In *Proceedings of the 21st International Conference on Artifi- cial Intelligence and Statistics (AISTATS)*, Vol. 84.

5. Gang Luo. 2016. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. *Health Information Science and Systems* 4, 2 (2016), 1–9. DOI: http://dx.doi.org/10.1186/s13755-016-0015-4

6. Thomas Mejtoft, Sarah Hale, and Ulrik Söderström. 2019. Design Friction: How intentionally added friction affect users' level of satisfaction. In *ECCE 2019*. 41–44. DOI:http://dx.doi.org/10.1145/3335082.3335106

7. Paul A. Pirraglia, Randall S. Stafford, and Daniel E. Singer. 2003. Trends in Prescribing of Selective Serotonin Reuptake Inhibitors and Other Newer Antidepressant Agents in Adult Primary Care. *The Primary Care Companion to The Journal of Clinical Psychiatry* 05, 04 (2003), 153–157. DOI: http://dx.doi.org/10.4088/pcc.v05n0402

8. Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

9. Melanie F. Pradier, Thomas H. McCoy Jr, Michael Hughes, Roy H. Perlis, and Finale Doshi-Velez. 2019. Predicting Treatment Discontinuation after Antidepressant Initiation. *Accepted to Translational Psychiatry* (2019).

10. Madhukar H. Trivedi and Ella J. Daly. 2008. Treatment strategies to improve and sustain remission in major depressive disorder. *Dialogues in Clinical Neuroscience* 10, 4 (2008), 377–384.

11. Q Yang, J Zimmerman, and A Steinfeld. 2015. Review of Medical Decision Support Tools: Emerging Opportunity for Interaction Design. *IASDR 2015 Interplay* September (2015), 1–16. DOI: http://dx.doi.org/10.13140/RG.2.1.1441.3284