Journal of Affective Disorders 306 (2022) 254–259

Contents lists available at ScienceDirect

Journal of Affective Disorders

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

Efficiently identifying individuals at high risk for treatment resistance in major depressive disorder using electronic health records

Isaac Lage [a], Thomas H. McCoy Jr [b,c], Roy H. Perlis [b,c,*], Finale Doshi-Velez [a,**]

[a] Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, 1 Oxford St, Science Center, 316.04, Cambridge, MA 02138, USA
[b] Center for Quantitative Health, Massachusetts General Hospital, 185 Cambridge Street, 6th Floor, Boston, MA 02114, USA
[c] Harvard Medical School, 25 Shattuck Street, Boston, MA 02115, USA

ARTICLE INFO

Keywords:
Treatment-resistant
Machine learning
Prediction
Risk stratification
Antidepressant
SSRI

ABSTRACT

Background: With the emergence of evidence-based treatments for treatment-resistant depression, strategies to identify individuals at greater risk for treatment resistance early in the course of illness could have clinical utility. We sought to develop and validate a model to predict treatment resistance in major depressive disorder using coded clinical data from the electronic health record.

Methods: We identified individuals from a large health system with a diagnosis of major depressive disorder receiving an index antidepressant prescription, and used a tree-based machine learning classifier to build a risk stratification model to identify those likely to experience treatment resistance. The resulting model was validated in a second health system.

Results: In the second health system, the extra trees model yielded an AUC of 0.652 (95% CI: 0.623–0.682); with sensitivity constrained at 0.80, specificity was 0.358 (95% CI: 0.300–0.413). Lift in the top quintile was 1.99 (95% CI: 1.76–2.22). Including additional data for the 4 weeks following treatment initiation did not meaningfully improve model performance.

Limitations: The extent to which these models generalize across additional health systems will require further investigation.

Conclusion: Electronic health records facilitated stratification of risk for treatment-resistant depression and demonstrated generalizability to a second health system. Efforts to improve upon such models using additional measures, and to understand their performance in real-world clinical settings, are warranted.

1. Introduction

The range of potential therapeutic options for treatment-resistant major depressive disorder (TRD) has broadened substantially beyond electroconvulsive therapy (ECT), for decades the gold standard in this setting (Pagnin et al., 2004; UK ECT Review Group, 2003). Recent developments include repetitive transcranial magnetic stimulation (rTMS), intravenous or intranasal ketamine, as well as atypical antipsychotic augmentation of standard antidepressants (Carter et al., 2020; Mantovani et al., 2012; Marcantoni et al., 2020). For the ~1/3 of individuals who do not benefit adequately from standard antidepressant treatments (Rush et al., 2006), these newer options offer potential opportunities to avoid the disability and chronicity typically associated with treatment resistance; however, each of these emerging strategies carry burdens which exceed those of standard first line antidepressants. As such, these more burdensome treatments are typically reserved for patients who have failed two – and in real-world settings often many more – standard treatments. Given recommendations about adequacy of acute treatment duration in depression (Kennedy et al., 2016), this would suggest a minimum of 6 months to establish treatment resistance, a period during

Abbreviations: TRD, Treatment-resistant major depressive disorder; ECT, Electroconvulsive therapy; rTMS, Repetitive transcranial magnetic stimulation; Major depressive disorder, MDD; ICD9, International Classification of Diseases Ninth Revision; ICD10, International Classification of Diseases Tenth Revision; ET, Extra trees classifier; LR, Logistic regression.

* Corresponding author at: Massachusetts General Hospital, 185 Cambridge Street, 6th Floor, Boston, MA 02114, USA.
** Corresponding author at: Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, 1 Oxford St, Science Center, 316.04, Cambridge, MA 02138, USA.
E-mail addresses: rperlis@mgh.harvard.edu (R.H. Perlis), finale@seas.harvard.edu (F. Doshi-Velez).

https://doi.org/10.1016/j.jad.2022.02.046
Received 19 November 2021; Received in revised form 7 February 2022; Accepted 14 February 2022
Available online 16 February 2022

which the intrinsic burden of depression symptoms and the varied morbidities of the syndrome persist.

If high-risk individuals could be identified earlier in their course, they might be prioritized for specialized evaluation and more intensive treatment (Simon and Perlis, 2010). At minimum, this could afford an opportunity to improve clinical outcomes. In addition, tools to stratify risk might enable more efficient investigation of novel therapeutics, by enabling enriched trials of interventions for treatment resistance in individuals most likely to benefit from them. To facilitate this identification, we utilized longitudinal data from the electronic health record of 2 large academic medical centers. We sought to apply standard machine learning methods to develop generalizable models to estimate risk for TRD among a large outpatient cohort, and validated these models in a large cohort from a second health system, as a starting point for more sophisticated approaches that might also incorporate biomarkers or other measures.

2. Methods

2.1. Study overview and cohort description

The study cohort included patients between 18 and 80 years old seen in the ambulatory networks of two academic medical centers in Eastern Massachusetts with an electronic prescription of one of 9 widely-used antidepressants (Table S1) and a diagnosis of major depressive disorder (MDD) or depressive disorder, not otherwise classified between March 2008 and December 2017. Diagnosis was defined as one or more instances of the International Classification of Diseases, Ninth Revision (ICD9) codes 296.2x, 296.3x, 311, or Tenth Revision (ICD10) codes f32.x and f33.x (Table S2). Patients with 2 or more schizophrenia or bipolar disorder diagnostic codes (Tables S3 and S4) were excluded for diagnostic ambiguity (Fig. S1).

Health records data for the relevant cohort were extracted to generate a data mart using i2b2 server software (i2b2, Boston, MA, USA) (Murphy et al., 2007). The data mart included sociodemographic features (age, gender, race and ethnicity), insurance information, diagnostic and procedure codes, inpatient medication administrations, and outpatient medication electronic prescriptions which were mandated during the period examined. The Mass General Brigham institutional review board approved the study protocol and waived the requirement for informed consent since this investigation used de-identified data and no human subjects contact was required.

2.2. Outcome definition

The primary study outcome was TRD, defined as 2 or more distinct antidepressant prescriptions in the first year after the index antidepressant (i.e., a total of at least 3 distinct antidepressants). We excluded individuals with any recorded prescriptions of antidepressants before March 2008, 6 months after electronic prescribing became mandatory in these hospital systems, to ensure we were observing index prescriptions. We also required that patients have at least 1 code at some points at least 12 weeks after their index antidepressant prescription. In a secondary analysis, we included features from the 4 weeks following the index prescription in addition to the features recorded in the 26 weeks before the index prescription, as a means of understanding whether available clinical data early in treatment course would improve outcome prediction.

2.3. Health records encoding and feature definitions

We encoded features for each patient from the 26-week window preceding the index antidepressant prescription; this threshold was selected a priori to balance the need for some prior features to facilitate prediction, with the desire to make predictions for as many individuals as possible. Features included sociodemographic variables, namely age, gender, and race; diagnostic codes, including both ICD-9 and ICD-10 codes; CPT codes capturing laboratory tests and procedures; and medications defined by Unified Medical Language System RxNorm codes (Bennett, 2012; "RxNorm," n.d.). All coded data were represented as count variables representing the number of times a feature was recorded in the specified time interval.

In a secondary analysis considering features from the 4 weeks following the index antidepressant prescription, we include the same features described above in both the 26-week window before index prescription, and in the 4-week window following it separately. That is, we included a feature for each code in each ontology (e.g. RxNorm, ICD) in the period before the index prescription, and in the 4 weeks following index antidepressant prescription.

A priori, we included all codes that occurred in 100 different patients in the feature time window in site A. For the main analysis, we selected 961 of the original 36,318 features. For the secondary analysis including the 4 weeks after index antidepressant prescription, we selected 1357 from 72,593 features, retaining separate features for codes recorded in the 26 weeks before the index prescription, and the 4 weeks after.

2.4. Prediction task

For model training, we randomly assigned the 22,006 eligible participants from site A to a training set (70%), validation set (10%) and testing set (20%) with labels stratified across the different sets of data (that is, each set has approximately the same number of participants with the relevant outcome). We repeated this 5 times, randomly splitting the patients into these sets each time, and trained a separate model for each of these training sets. (Repeating these analyses sampling 10 times did not yield meaningfully different results.) Our reported results are averaged across the 5 models and accompanying test sets, unless otherwise specified. All site B participants (12,869) were held out to test our generalization performance. (For the secondary analysis including data from 4 weeks after index prescription, site A had 21,774 patients, and site B had 12,755 after excluding participants experiencing outcome in the 4-week window following index antidepressant prescription).

2.5. Classification methods and metrics

We trained a commonly-used classifier, extra trees (ET), and for comparison, a logistic regression (LR) model. For both, we used the open-source Scikit Learn toolkit implementation (Pedregosa et al., 2011). We tuned hyperparameters to achieve the best AUC on the site A validation set using grid search over the options specified below. For each of the 5 random train/valid/test splits of the dataset, the hyperparameter search was repeated to choose the best hyperparameter setting for that train/validation set.

For LR, we used l2 regularization, and we searched over 1 hyperparameter: the regularization strength, denoted "C", (C={20 points between 1e-5 and 1 spaced linearly on a log-base-10 scale}). For ET, we searched over three hyperparameters: the fraction of features used in each split (max_features={0.04, 0.16, 0.64}), the minimum number of samples at leaf nodes (min_samples_leaf={16, 64, 256}), and the number of estimators (n_estimators={16, 64, 256}). (The regularization chosen for each trained model in each analysis is listed in Table S5.) For both models, features were z-scored so as to have similar magnitudes for all dimensions, and we set the class labels to have equal weight in the optimization.

We reported model performance using AUC, and specificity, PPV and NPV with the threshold chosen to set sensitivity as close to 0.80 as possible (i.e., to define a model with 80% sensitivity for treatment-resistance). We computed these metrics on the independent dataset from site B, as well as the held-out test set from site A. We averaged results over 5 random train/valid/test splits of the site A data (for which a model was trained for each), and we then took 500 bootstrap samples of the points in each test set to compute the mean and the 95%

confidence intervals (i.e., scores at 2.5 and 97.5 percentiles) for each metric. We evaluated the 5 models trained on site A on site B using the same bootstrapping procedure. A separate threshold was chosen for each of the bootstrap samples when computing specificity, PPV and NPV.

3. Results

Characteristics of the cohort in the two sites are summarized in Table 1. A total of 1079 individuals out of 22,006 individuals (4.90%) in the site A cohort were prescribed more than 2 antidepressants within the first year of the index prescription, and 520/12,869 individuals (4.04%) among those from Site B. Sociodemographic features including age, sex, and race/ethnicity were similar between those with and without TRD at both sites. Rates of TRD were greatest among individuals receiving mirtazapine or duloxetine (Table S1) as index antidepressant at site A (14% and 11%, respectively), while in site B, the TRD outcome was more evenly distributed across treatments, ranging from 4% to 8%.

Table 2 reports classifier performance metrics for the primary analysis, predicting TRD based on features drawn from the 26 preceding weeks (6 months) of clinical data. Metrics for extra trees are reported both on an independent testing sample from Site A, and the full Site B cohort, as well as for both classes of models (Table 2; The most predictive features in both ET and LR models are presented in Table S6). For the primary analysis, the ET model achieved an AUC of 0.652 (0.623 – 0.682) in site B; constraining sensitivity to be as close as possible to 0.80, specificity was 0.358 (0.300–0.413), NPV 0.977 (0.972–0.981), and PPV 0.050 (0.044–0.056). By comparison, performance was somewhat poorer using logistic regression (Table 2), with an AUC of 0.614 (0.591 - 0.637). (For individual features and model weights or coefficients, please see Supplemental Table 6.) Incorporating an additional 4 weeks after initial prescription did not meaningfully improve prediction (Table 2).

We next examined AUC for the primary ET model stratified by gender and race (Table 3) to establish performance in subsamples of the cohort. The AUC in site B for white patients (0.657; 0.621–0.695) and non-white patients (0.639; 0.593–0.684) was similar, as were AUCs for female patients (0.654; 0.621–0.686) and male patients (0.647; 0.590–0.704). Other characteristics of discrimination were likewise similar between subgroups.

Finally, we examined calibration in the ET model Fig. 1 illustrates the rate of TRD for each risk quintile for the main analysis, computed on the basis of predicted probabilities. The horizontal dashed line represents the mean risk of developing the outcome. The lift in the top quintile was 1.99 (1.76–2.22) for site B, and 2.12 (1.73–2.50) for the site A testing set with the ET model.

4. Discussion

In this investigation of 34,877 individuals with a diagnosis of MDD initiated on standard antidepressant treatment, we developed models able to stratify risk for treatment-resistance. Performance of these models is modest compared to prediction tools in other areas of medicine (Castro et al., 2020), but similar in magnitude to other treatment outcome prediction models relating to antidepressant treatment. Among the specific features associated with greater risk in both models are initial use of non-SSRI antidepressants; rather than reflecting differential efficacy per se, this result likely suggests other patient characteristics such as comorbidity that might lead the clinician to select alternate treatment. The inclusion of clinical features from early in treatment, which we hypothesized might improve predictions by giving indications of propensity to tolerate medications and/or placebo-like response, did not meaningfully improve these predictions.

Since one of us published the first machine learning model of treatment-resistant depression using the STAR*D data set (Perlis, 2013), multiple additional studies have used those data, alone or in concert with additional clinical trials, to try to improve performance (Iniesta et al., 2016; Nie et al., 2018). Such clinical trial-based studies generally face three major challenges: they use research measures that would be challenging to deploy at scale in clinical settings; they are prone to overfit when they incorporate large numbers of features relative to the number of observations; and they are unlikely to reflect real-world clinical populations. Subsequent efforts using clinical cohort studies also reported promising results, but still rely on systematic/structured assessment and have unclear generalizability (Kautzky et al., 2017). In general, despite a proliferation of machine-learning efforts, larger-scale and more systematic investigations tend to yield less optimistic estimates of discrimination.

The present results are not directly comparable to prior work in individual clinical trials, as this study relies solely on artifacts of routine clinical care – i.e., secondary use of data available in the electronic health record at time of prescription. In particular, rates of TRD are far lower than the 1 in 3 commonly cited based upon the STAR*D effectiveness study (Rush et al., 2006). Perhaps unsurprisingly, the resulting discrimination is also lower in absolute terms than that previously reported using STAR*D data (Perlis, 2013). On the other hand, by examining model characteristics in a cohort drawn from a second health system, these estimates are likely to represent a pragmatic estimate of how claims code-based models will perform in real-world settings.

Table 1
Cohort sociodemographic data at site A and site B.

	Count/Mean full sample	%/Std dev. full sample	Count/Mean TRD sample	%/Std dev. TRD sample
Site A	22,006	–	–	–
TRD Prevalence	1079	4.90	–	–
Gender: F	14,531	66.0	718	66.5
Gender: M	7475	34.0	361	33.5
Race: Asian	575	2.6	26	2.4
Race: Black	995	4.5	46	4.3
Race: Hispanic	1226	5.6	52	4.8
Race: Other	1303	5.9	68	6.3
Race: White	17,907	81.4	887	82.2
Age	47.3	14.9	46.3	14.5
Site B	12,869	–	–	–
TRD Prevalence	520	4.04	–	–
Gender: F	9245	71.8	375	72.1
Gender: M	3624	28.2	145	27.9
Race: Asian	207	1.6	11	2.1
Race: Black	1073	8.3	38	7.3
Race: Hispanic	1533	11.9	75	14.4
Race: Other	721	5.6	34	6.5
Race: White	9335	72.5	362	69.6
Age	48.1	14.5	46.4	14.2

Table 2
Classifier performance metrics for both models in each site in the main analysis and the secondary analysis. PPV and NPV are reported for the threshold setting sensitivity as close as possible to 0.80.

Analysis	Model	Site	AUC	AUC 95% CI	Specificity	Specificity 95% CI	PPV	PPV 95% CI	NPV	NPV 95% CI
Main Analysis	Logistic Regression	A	0.646	0.602 - 0.693	0.349	0.254 - 0.435	0.060	0.050 - 0.071	0.971	0.961 - 0.978
		B	0.614	0.586 - 0.642	0.315	0.241 - 0.377	0.047	0.041 - 0.053	0.974	0.966 - 0.979
	Extra Trees	A	0.666	0.609 - 0.723	0.382	0.253 - 0.529	0.063	0.050 - 0.081	0.973	0.960 - 0.982
		B	0.652	0.623 - 0.682	0.358	0.300 - 0.413	0.050	0.044 - 0.056	0.977	0.972 - 0.981
Secondary Analysis: + 4 weeks after	Logistic Regression	A	0.650	0.595 - 0.701	0.374	0.262 - 0.486	0.050	0.039 - 0.063	0.978	0.969 - 0.985
		B	0.631	0.598 - 0.661	0.353	0.287 - 0.407	0.039	0.034 - 0.044	0.982	0.977 - 0.985
	Extra Trees	A	0.659	0.606 - 0.715	0.407	0.314 - 0.515	0.052	0.042 - 0.065	0.980	0.974 - 0.985
		B	0.648	0.603 - 0.682	0.378	0.300 - 0.461	0.041	0.035 - 0.048	0.983	0.978 - 0.986

Table 3
Classifier performance metrics stratified by race and gender for the extra trees and the logistic regression model in the main analysis.

Model	Site	Subgroup	AUC	AUC 95% CI	Specificity	Specificity 95% CI	PPV	PPV 95% CI	NPV	NPV 95% CI
Logistic Regression	A	White	0.646	0.600 - 0.691	0.347	0.217 - 0.459	0.061	0.047 - 0.076	0.970	0.956 - 0.979
		NonWhite	0.655	0.536 - 0.773	0.361	0.115 - 0.568	0.058	0.026 - 0.099	0.972	0.944 - 0.987
		Male	0.644	0.571 - 0.718	0.350	0.180 - 0.488	0.061	0.036 - 0.085	0.970	0.952 - 0.982
		Female	0.650	0.590 - 0.709	0.346	0.197 - 0.490	0.060	0.045 - 0.078	0.970	0.952 - 0.981
	B	White	0.614	0.582 - 0.645	0.313	0.229 - 0.381	0.045	0.038 - 0.052	0.974	0.966 - 0.980
		NonWhite	0.615	0.568 - 0.661	0.317	0.171 - 0.437	0.052	0.041 - 0.065	0.970	0.948 - 0.980
		Male	0.597	0.543 - 0.646	0.304	0.199 - 0.421	0.046	0.036 - 0.058	0.972	0.957 - 0.982
		Female	0.622	0.587 - 0.654	0.316	0.231 - 0.389	0.047	0.041 - 0.054	0.974	0.964 - 0.979
Extra Trees	A	White	0.661	0.607 - 0.720	0.385	0.247 - 0.534	0.065	0.049 - 0.087	0.972	0.960 - 0.981
		NonWhite	0.675	0.527 - 0.802	0.311	0.101 - 0.600	0.055	0.025 - 0.101	0.964	0.916 - 0.986
		Male	0.662	0.580 - 0.766	0.375	0.169 - 0.585	0.064	0.038 - 0.093	0.971	0.940 - 0.985
		Female	0.671	0.597 - 0.730	0.381	0.233 - 0.511	0.063	0.046 - 0.085	0.973	0.957 - 0.982
	B	White	0.657	0.621 - 0.695	0.355	0.286 - 0.430	0.048	0.041 - 0.055	0.978	0.972 - 0.982
		NonWhite	0.639	0.593 - 0.684	0.342	0.220 - 0.433	0.054	0.044 - 0.066	0.973	0.961 - 0.981
		Male	0.647	0.590 - 0.704	0.350	0.230 - 0.453	0.049	0.039 - 0.061	0.976	0.965 - 0.983
		Female	0.654	0.621 - 0.686	0.359	0.287 - 0.425	0.050	0.043 - 0.058	0.977	0.971 - 0.981

Beyond the advantage of generalizability, we find that the performance metrics we measure are not markedly different across sociodemographic subgroups, suggesting less evidence of bias than some other machine learning models, although this remains an important consideration for future work. (One notable exception is the finding of greater rates of treatment resistance among individuals in one health system beginning non-SSRI antidepressants; this may reflect differences in care systems rather than biology per se, with some clinical settings adopting particular antidepressants in higher-risk or more comorbid patient populations.)

In interpreting our results, it is important to consider the limitations inherent in electronic health records. First, we rely on recorded prescriptions, not medication fills, so cannot estimate adherence, which may contribute substantially to apparent treatment resistance. In prior work examining antidepressant blood levels in discarded samples, we found detectable antidepressant levels in more than 80% of participants with electronic prescriptions (Roberson et al., 2016). Second, we cannot exclude antidepressant prescriptions or other treatment received outside of these health systems; as such, it is possible that some individuals have greater treatment resistance than we detect, or that some of those identified as responsive were misclassified. Similarly, although we exclude individuals with antidepressant treatment in the 26 weeks prior to index prescription, in order to maximize the likelihood of studying treatment-naïve individuals, some patients undoubtedly received past treatment. In aggregate, all of these factors should tend to diminish model performance and may impact portability, but are unlikely to introduce systematic bias.

We note that, as with any study using naturalistic data, numerous a priori decisions are required in defining predictors and outcomes. For example, we exclude ICD9 code 300.4 (dysthymia) on the basis of prior observations that the diagnosis may be less reliable (Castro-Rodríguez et al., 2015) and that these individuals are less likely to be treated with antidepressants, while retaining 311 as it is widely used in primary care settings and included in many code-based studies of depression (Simon et al., 2015;Simon and Savarino, 2008;Bobo et al., 2020;Adekkanattu et al., 2018;Pilon et al., 2019;Nguyen et al., 2008). While consistent with prior definitions, the use of diagnostic codes beyond those strictly reflecting major depressive disorder could impact validity of these results. We also limited predictors to facts observable 26 weeks prior to initial prescription, in an effort to balance inclusivity (i.e., being able to make predictions for as many individuals as possible, favoring shorter lead-in) with the need for at least some facts in order to generate predictions in the first place (i.e., having sufficient lead-in to observe facts that might impact outcome, favoring longer lead-in); the optimal window depends entirely on how and where such models might be applied and is likely to differ between care settings. We examined 12 month follow-up to minimize loss to follow-up; prior work examines models for predicting discontinuation after treatment initiation (Pradier et al., 2020). In particular, we cannot distinguish when one episode ends and another begins, reflecting the more fluctuating and chronic course of illness often observed in real-world settings (Perlis et al., 2012). And, while definitions of treatment-resistance vary, we selected the need for a 3rd antidepressant (i.e., 2 prior treatments without remission) to reflect the most common. Altering any of these to fit alternate clinical

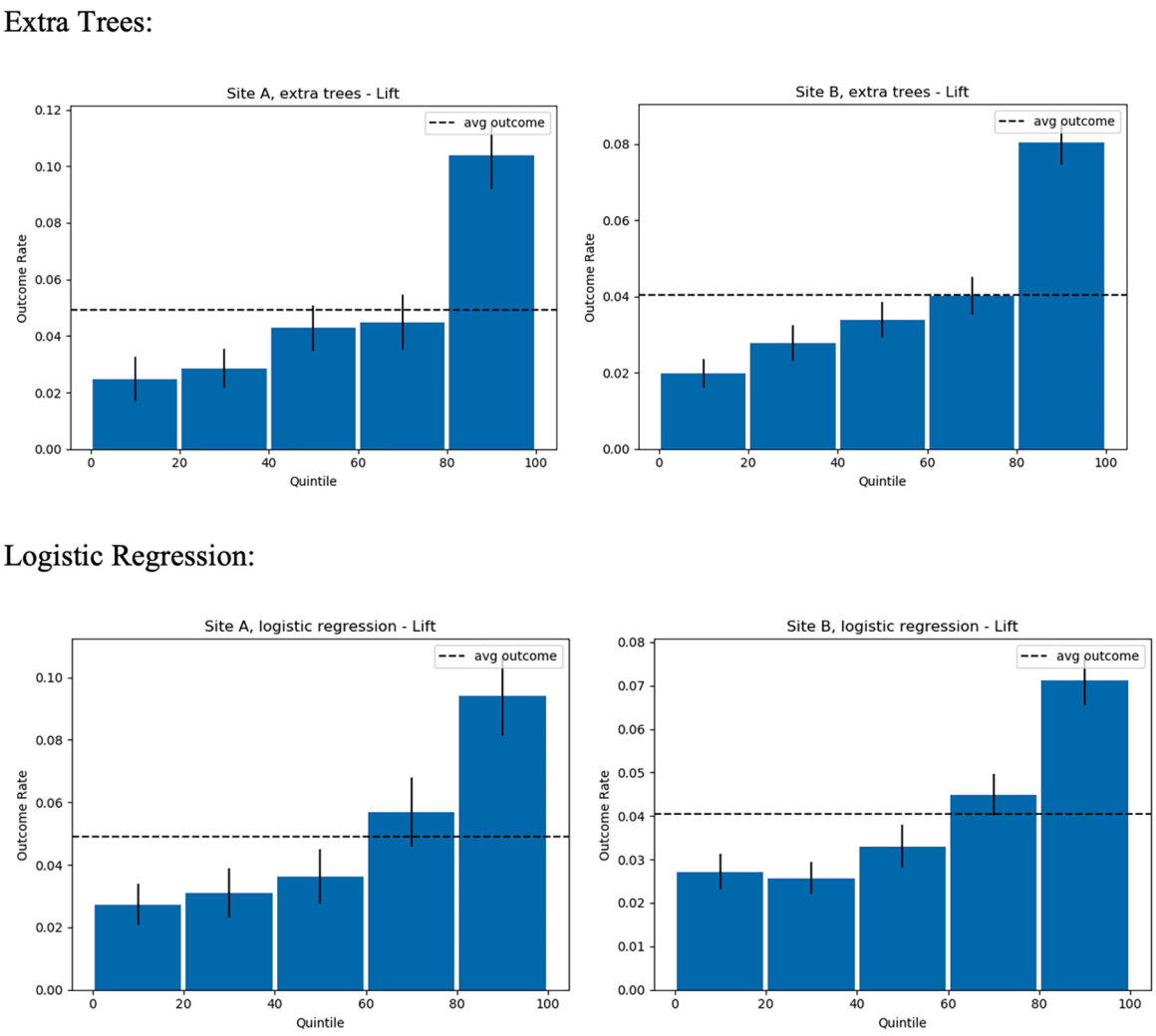

Fig. 1. Outcome rate for each quintile by predicted risk for the main analysis in each site with the extra tree and logistic regression models.
Legend: The dotted black line shows the mean outcome across all quintiles. The error bars show the standard deviations. From this, lift can be ascertained.

use-cases, such as a desire to make predictions with less lead-in, or to predict lesser degrees of treatment resistance, would be reasonable strategies for future work. Likewise, whether the metrics we report suggest sufficient performance to warrant clinical application depends entirely on clinical context; there is no 'magic threshold' at which a test is good enough (Perlis, 2011; Uher et al., 2012).

Incorporation of additional measures could improve prediction substantially: the use of research measures appears to yield more performant risk predictions (Perlis, 2013), and incorporation of features derived from clinical free text also improves prediction (McCoy et al., 2018, 2016), although free text-based models may be more challenging to transfer from one health system to another. In fact, one application of these models might be to identify higher-risk individuals who then receive more comprehensive assessment to derive more reliable predictions (Perlis et al., 2018). Particularly for biomarkers such as neuroimaging (Drysdale et al., 2017), clinical risk stratification might represent a first step to enrich for individuals more likely to benefit from intensive assessment. Integration of electronic health record data with such assessments, or with other data types such as genomic data, merits further study. Particularly for higher-cost or more intensive interventions, a stepped approach to assessment still offers likely advantages over the trial-and-error approach (Maxfield and Zineh, 2021) to treatment resistance currently applied.

Data availability

Our institutional review has indicated that these data cannot be redistributed.

CRediT authorship contribution statement

Isaac Lage: Formal analysis, Writing – original draft, Writing – review & editing. **Thomas H. McCoy Jr:** Writing – original draft, Writing – review & editing. **Roy H. Perlis:** Project administration, Writing – original draft, Writing – review & editing. **Finale Doshi-Velez:** Project administration, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

THM receives research funding from the Brain and Behavior Research Foundation, Telefonica Alfa, National Institute of Mental Health, National Institute of Nursing Research and National Library of Medicine. FDV consults for Davita Kidney Care and Google Health. RHP holds equity in Psy Therapeutics and Outermost Therapeutics; serves on the scientific advisory boards of Genomind and Takeda; and consults to RID Ventures. RHP receives research funding from NIMH, NHLBI, NHGRI, and Telefonica Alfa. The other authors have no disclosures to report.

Acknowledgments

This work was funded by Oracle Labs, Harvard SEAS, the Blyth Family Fund, the Harvard Data Science Initiative, the NSF GRFP (grant no. DGE1745303; IL), and a grant from the National Institute of Mental Health (grant no. 1R01MH106577; RHP). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

The authors thank Victor Castro for assistance with dataset generation.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jad.2022.02.046.

References

Adekkanattu, P., Sholle, E.T., DeFerio, J., Pathak, J., Johnson, S.B., Campion, T.R., 2018. Ascertaining depression severity by extracting patient health questionnaire-9 (PHQ-9) scores from clinical notes. AMIA Annu. Symp. Proc. 2018, 147–156.

Bennett, C.C., 2012. Utilizing RxNorm to support practical computing applications: capturing medication history in live electronic health records. J. Biomed. Inform. 45, 634–641. https://doi.org/10.1016/j.jbi.2012.02.011.

Bobo, W.V., Ryu, E., Petterson, T.M., Lackore, K., Cheng, Y., Liu, H., Suarez, L., Preisig, M., Cooper, L.T., Roger, V.L., Pathak, J., Chamberlain, A.M., 2020. Bi-directional association between depression and HF: an electronic health records-based cohort study. J. Comorbidity 10. https://doi.org/10.1177/2235042X20984059, 2235042X20984059.

Carter, B., Strawbridge, R., Husain, M.I., Jones, B.D.M., Short, R., Cleare, A.J., Tsapekos, D., Patrick, F., Marwood, L., Taylor, R.W., Mantingh, T., de Angel, V., Nikolova, V.L., Carvalho, A.F., Young, A.H., 2020. Relative effectiveness of augmentation treatments for treatment-resistant depression: a systematic review and network meta-analysis. Int. Rev. Psychiatry Abingdon Engl. 32, 477–490. https://doi.org/10.1080/09540261.2020.1765748.

Castro, V.M., McCoy, T.H., Perlis, R.H., 2020. Laboratory findings associated with severe illness and mortality among hospitalized individuals with coronavirus disease 2019 in Eastern Massachusetts. JAMA Netw. Open 3, e2023934. https://doi.org/10.1001/jamanetworkopen.2020.23934.

Castro-Rodríguez, J.I., Olariu, E., Garnier-Lacueva, C., Martín-López, L.M., Pérez-Solà, V., Alonso, J., Forero, C.G., INSAyD Investigators, 2015. Diagnostic accuracy and adequacy of treatment of depressive and anxiety disorders: a comparison of primary care and specialized care patients. J. Affect. Disord. 172, 462–471. https://doi.org/10.1016/j.jad.2014.10.020.

Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R. N., Zebley, B., Oathes, D.J., Etkin, A., Schatzberg, A.F., Sudheimer, K., Keller, J., Mayberg, H.S., Gunning, F.M., Alexopoulos, G.S., Fox, M.D., Pascual-Leone, A., Voss, H.U., Casey, B.J., Dubin, M.J., Liston, C., 2017. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat. Med. 23, 28–38. https://doi.org/10.1038/nm.4246.

Iniesta, R., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., Henigsberg, N., Dernovsek, M.Z., Souery, D., Stahl, D., Dobson, R., Aitchison, K.J., Farmer, A., Lewis, C.M., McGuffin, P., Uher, R., 2016. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. J. Psychiatr. Res. 78, 94–102. https://doi.org/10.1016/j.jpsychires.2016.03.016.

Kautzky, A., Baldinger-Melich, P., Kranz, G.S., Vanicek, T., Souery, D., Montgomery, S., Mendlewicz, J., Zohar, J., Serretti, A., Lanzenberger, R., Kasper, S., 2017. A new prediction model for evaluating treatment-resistant depression. J. Clin. Psychiatry 78, 215–222. https://doi.org/10.4088/JCP.15m10381.

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., CANMAT Depression Work Group, 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. Psychiatry Rev. Can. Psychiatr. 61, 540–560. https://doi.org/10.1177/0706743716659417.

Mantovani, A., Pavlicova, M., Avery, D., Nahas, Z., McDonald, W.M., Wajdik, C.D., Holtzheimer, P.E., George, M.S., Sackeim, H.A., Lisanby, S.H., 2012. Long-term efficacy of repeated daily prefrontal transcranial magnetic stimulation (TMS) in treatment-resistant depression. Depress. Anxiety 29, 883–890. https://doi.org/10.1002/da.21967.

Marcantoni, W.S., Akoumba, B.S., Wassef, M., Mayrand, J., Lai, H., Richard-Devantoy, S., Beauchamp, S., 2020. A systematic review and meta-analysis of the efficacy of intravenous ketamine infusion for treatment resistant depression: January 2009 - January 2019. J. Affect. Disord. 277, 831–841. https://doi.org/10.1016/j.jad.2020.09.007.

Maxfield, K., Zineh, I., 2021. Precision dosing: a clinical and public health imperative. JAMA. https://doi.org/10.1001/jama.2021.1004.

McCoy, T.H., Castro, V.M., Roberson, A.M., Snapper, L.A., Perlis, R.H., 2016. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. JAMA Psychiatry 73, 1064. https://doi.org/10.1001/jamapsychiatry.2016.2172.

McCoy, T.H., Yu, S., Hart, K.L., Castro, V.M., Brown, H.E., Rosenquist, J.N., Doyle, A.E., Vuijk, P.J., Cai, T., Perlis, R.H., 2018. High throughput phenotyping for dimensional psychopathology in electronic health records. Biol. Psychiatry 83, 997–1004. https://doi.org/10.1016/j.biopsych.2018.01.011.

Murphy, S.N., Mendis, M., Hackett, K., Kuttan, R., Pan, W., Phillips, L.C., Gainer, V., Berkowicz, D., Glaser, J.P., Kohane, I., Chueh, H.C., 2007. Architecture of the open-source clinical research chart from informatics for integrating biology and the bedside. AMIA. Annu. Symp. Proc. 548–552.

Nguyen, H.Q., Koepsell, T., Unützer, J., Larson, E., LoGerfo, J.P., 2008. Depression and use of a health plan–sponsored physical activity program by older adults. Am. J. Prev. Med. 35, 111–117. https://doi.org/10.1016/j.amepre.2008.04.014.

Nie, Z., Vairavan, S., Narayan, V.A., Ye, J., Li, Q.S., 2018. Predictive modeling of treatment resistant depression using data from STAR*D and an independent clinical study. PLoS One 13. https://doi.org/10.1371/journal.pone.0197268.

Pagnin, D., de Queiroz, V., Pini, S., Cassano, G.B., 2004. Efficacy of ECT in depression: a meta-analytic review. J. ECT 20, 13–20. https://doi.org/10.1097/00124509-200403000-00004.

Pedregosa, F., Veroquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Perlis, R.H., 2013. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. Biol. Psychiatry 74, 7–14. https://doi.org/10.1016/j.biopsych.2012.12.007.

Perlis, R.H., 2011. Translating biomarkers to clinical practice. Mol. Psychiatry 16, 1076–1087. https://doi.org/10.1038/mp.2011.63.

Perlis, R.H., Fava, M., McCoy Jr, T.H., 2018. Can electronic health records revive central nervous system clinical trials? Mol. Psychiatry. https://doi.org/10.1038/s41380-018-0278-z [Epub ahead of print].

Perlis, R.H., Iosifescu, D.V., Castro, V.M., Murphy, S.N., Gainer, V.S., Minnier, J., Cai, T., Goryachev, S., Zeng, Q., Gallagher, P.J., Fava, M., Weilburg, J.B., Churchill, S.E., Kohane, I.S., Smoller, J.W., 2012. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol. Med. 42, 41–50. https://doi.org/10.1017/S0033291711000997.

Pilon, D., Joshi, K., Sheehan, J.J., Zichlin, M.L., Zuckerman, P., Lefebvre, P., Greenberg, P.E., 2019. Burden of treatment-resistant depression in medicare: a retrospective claims database analysis. PLoS One 14, e0223255. https://doi.org/10.1371/journal.pone.0223255.

Pradier, M.F., McCoy, T.H., Hughes, M., Perlis, R.H., Doshi-Velez, F., 2020. Predicting treatment dropout after antidepressant initiation. Transl. Psychiatry 10, 60. https://doi.org/10.1038/s41398-020-0716-y.

Roberson, A.M., Castro, V.M., Cagan, A., Perlis, R.H., 2016. Antidepressant nonadherence in routine clinical settings determined from discarded blood samples. J. Clin. Psychiatry 77, 359–362. https://doi.org/10.4088/JCP.14m09612.

Rush, A.J., Trivedi, M.H., Wisniewski, S.R., Nierenberg, A.A., Stewart, J.W., Warden, D., Niederehe, G., Thase, M.E., Lavori, P.W., Lebowitz, B.D., McGrath, P.J., Rosenbaum, J.F., Sackeim, H.A., Kupfer, D.J., Luther, J., Fava, M., 2006. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. Am. J. Psychiatry 163, 1905–1917. https://doi.org/10.1176/ajp.2006.163.11.1905.

Simon, G.E., Perlis, R.H., 2010. Personalized medicine for depression: can we match patients with treatments? Am. J. Psychiatry 167, 1445–1455. https://doi.org/10.1176/appi.ajp.2010.09111680.

Simon, G.E., Rossom, R.C., Beck, A., Waitzfelder, B.E., Coleman, K.J., Stewart, C., Operskalski, B., Penfold, R.B., Shortreed, S.M., 2015. Antidepressants are not overprescribed for mild depression. J. Clin. Psychiatry 76, 1627–1632. https://doi.org/10.4088/JCP.14m09162.

Simon, G.E., Savarino, J., 2008. Suicide attempts among patients starting depression treatment with medications or psychotherapy. Focus 6, 80–85. https://doi.org/10.1176/foc.6.1.foc80 (Madison).

Uher, R., Tansey, K.E., Malki, K., Perlis, R.H., 2012. Biomarkers predicting treatment outcome in depression: what is clinically significant? Pharmacogenomics 13, 233–240. https://doi.org/10.2217/pgs.11.161.

UK ECT Review Group, 2003. Efficacy and safety of electroconvulsive therapy in depressive disorders: a systematic review and meta-analysis. Lancet Lond. Engl. 361, 799–808. https://doi.org/10.1016/S0140-6736(03)12705-5.