Title: Generating interpretable predictions about antidepressant treatment stability using supervised topic models

Authors: Michael C. Hughes, PhD[1]; Melanie F. Pradier, PhD[2]; Andrew Slavin Ross, MEng[2]; Thomas H. McCoy Jr, MD[3, 4]; Roy H. Perlis, MD MSc[3, 4]; Finale Doshi-Velez, PhD[2]

Affiliations

[1] Department of Computer Science, Tufts University, Medford, MA
[2] John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA
[3] Center for Quantitative Health, Massachusetts General Hospital, Boston, MA
[4] Harvard Medical School, Boston, MA

Corresponding Author

Michael C. Hughes, PhD
161 College Avenue
Medford, MA 02155
mhughes@cs.tufts.edu

Manuscript Contents

Links

Supplement document providing more results, links to interactive visualizations, and detailed procedures for reproducibility
https://www.michaelchughes.com/papers/HughesEtAl_medRxiv2020_Supplement.pdf

STROBE checklist:
https://www.michaelchughes.com/papers/HughesEtAl_medRxiv2020_STROBE_checklist.pdf

Open-source code for our proposed machine learning methods
https://github.com/dtak/prediction-constrained-topic-models

Key Points

Question: How well can coded clinical data from electronic health records be used to predict achievement of a stable antidepressant regimen in major depressive disorder?

Findings: In this in silico cohort study of 81,630 adults, we identified 55,303 who reached a stable antidepressant treatment regimen. Predictions using generalized linear models or ensembles of decision trees applied to diagnosis, procedure, and medication codes, as well as low-dimensional summaries of these codes via supervised topic models, achieved area under receiver operating characteristic curve values of ~0.62-0.65; treatment-specific models performed no better than general treatment outcome models.

Meaning: Coded clinical data can facilitate prediction of antidepressant treatment outcomes, but medication-specific models do not outperform general response prediction models.

Abstract

Importance: In the absence of readily-assessed and clinically-validated predictors of treatment response, pharmacologic management of major depressive disorder (MDD) often relies on trial and error.

Objective: To utilize electronic health records to identify predictors of treatment response, while preserving interpretability of predictions despite large numbers of covariates.

Design: Retrospective cohort study.

Setting: Two academic medical centers in Boston, including outpatient primary and specialty care clinics.

Participants: 81,630 adults with a coded diagnosis of MDD.

Exposure: Treatment with 1 or more of 11 standard antidepressants.

Main Outcomes and Methods: Stable treatment, intended as a proxy for treatment effectiveness, defined as continued prescription of an antidepressant for 90 days. We trained supervised topic models to extract 10 interpretable covariates from coded clinical data for stability prediction. Then, using data from one hospital system (Site A) we trained generalized linear models and ensembles of decision trees to predict stability outcomes from topic features that summarize patient history. We evaluated on held-out patients from Site A as well as all individuals from a second hospital system (B).

Results: Among the 81,630 adults (31% male; age 18-80 with mean 48.46), we identified 55,303 who reached a stable treatment regimen during follow-up. For held-out patients from Site A, mean area-under-the-receiver-operating-characteristic-curve (AUC) discrimination for general stability outcome was 0.627 (95% confidence interval (CI) 0.615 - 0.639) for our supervised topic model with 10 covariates. In evaluation on site B, our approach achieved similar AUC of 0.619 (95% CI 0.610 - 0.627). Building models to predict stability *specific* to a particular drug did not improve upon predicting general stability, even when using a harder-to-interpret ensemble classifier and 9,256 coded covariates (specific AUC = 0.647, 95% CI 0.635- 0.658; general AUC = 0.661, 95% CI 0.648-0.672). Topics coherently captured clinical concepts associated with treatment response.

Conclusions and Relevance: Coded clinical data available in electronic health records facilitated prediction of general treatment response, but not response to specific medications. While greater discrimination is likely required for clinical application, our results provide a simple and transparent baseline for such studies.

Funding: Oracle Labs, Harvard SEAS, and National Institute of Mental Health.

Introduction

Meta-analysis suggests that newer antidepressants are on average very similar in efficacy and overall tolerability[1] a finding further supported by a small number of effectiveness studies[2–4]. However, these group averages obscure a wide amount of interindividual variability; even before recent enthusiasm for precision or personalized medicine, a vast literature addressed potential predictors of antidepressant treatment outcome aimed at identifying individuals who are more or less likely to benefit[5]. For example, symptom-defined subtypes were investigated, initially as predictors of tricyclic antidepressant or monoamine oxidase inhibitor response, then as predictors of selective serotonin reuptake inhibitor response[6–8]. More recently, rather than clinical subtypes, efforts have focused on deriving constellations of symptoms more predictive of response[9–11] or incorporating additional survey measures[12]. Beyond clinical predictors, numerous studies examined incorporation of biomarkers, most notably (and notoriously) the dexamethasone suppression test[13,14].

A key challenge in all of these studies has been the paucity of head-to-head antidepressant studies, such that it is often difficult to distinguish predictors of poor outcomes overall from predictors of poor outcomes specific to a given medication. Traditional tests of interaction compound this problem, as they are best-powered for opposing effects (i.e., markers associated with better outcome in one group and poorer outcome in another), when in reality this may not comport with biology. Further, even where head-to-head studies exist, there are rarely replication cohorts to follow up initial associations.

In other contexts, electronic health record (EHR) or administrative data sets have been used to predict clinical outcomes, providing sufficiently large real-world cohorts to allow identification and validation of predictors[15–17]. They may offer the further advantage of operating on data already readily available at the point of care, such that clinical adoption does not require the use of new rating scales or measures. In the present study, we sought to apply widely available health record data to understand two questions: first, to what extent can we identify general (i.e. non-specific) predictors of antidepressant response? Second, can we identify treatment-specific predictors, as might be applied to drive a precision medicine approach to antidepressant prescribing?

In so doing, we also investigated a potential solution to a central problem in analysis of large clinical datasets and machine learning for big data more generally: a lack of interpretability[18–20]. While optimized predictions may be useful, the inability to understand what drives these predictions may impede efforts to validate and disseminate them in clinical settings. Moreover, the reliance on individual clinical data points may limit portability if health systems use different procedure or diagnostic codes to reflect the same underlying concepts. Here, we applied a recently-developed supervised topic modeling approach[21] that yields much simpler predictors based on groups of features that retain discrimination and facilitate interpretability.

Methods

Study Design

We utilized an in silico cohort drawn from EHR to examine the association between coded EHR available at time of medication prescription for standard antidepressants and subsequent longitudinal outcome of stable treatment with that medication.

The study cohort included individuals with at least one diagnosis of major depressive disorder

(MDD, ICD9 codes 296.2x, 296.3x) or depressive disorder not otherwise specified (ICD9 code 311) who received psychiatric care between 1997 and 2017 across the inpatient and outpatient networks of two large academic medical centers in New England ("Site A" and "Site B"). Patients were excluded if age was less than 18 or greater than 80, if the total observation period was less than 90 days, or if there were fewer than 3 total documented visits (of any type, psychiatric or otherwise) in the EHR.

We extracted de-identified patient-level data using the i2b2 server software (i2b2, Boston, MA, USA)[22]. Available patient data included socio-demographic information (age, sex, race/ethnicity), all diagnostic and procedure codes, and all inpatient and outpatient medication prescriptions.

After applying inclusion criteria (Supplemental eFigure 1), a total of 51,048 patients from Site A were included and randomly assigned to training (50%, n=25524), validation (25%, n=12762), and test (25%, n=12762) subsets. 26,176 patients from Site B form an external validation set.

The Partners HealthCare institutional review board approved the study protocol, waiving the requirement for informed consent as only de-identified data was utilized and no human subjects contact was required.

Outcome Definition

Recognizing that traditional clinical trial outcomes such as response and remission are difficult to define reliably for all individuals using solely coded clinical data[16], we instead sought to identify individuals who achieved a period of stable treatment, as a proxy for ample clinical benefit and tolerability. We applied a simplifying but face-valid assumption that successful treatments continue uninterrupted over time with repeated prescriptions, while unsuccessful treatments are either discontinued or require addition of further medication[4].

We initially considered 27 possible antidepressants (see Supplemental eTable 1). We defined a treatment segment as stable if it contained at least two prescriptions for the same antidepressant(s) on two distinct dates, the total duration was at least 90 days, the calculated medication possession ratio (fraction of days in segment where the patient possessed a valid, non-expired prescription)[23] was at least 80%, and the largest gap between start dates of any two prescriptions in the segment was between 30 days and 390 days. See Supplemental eFigures 2 and 3 for illustrative examples. Only 11 antidepressants had sufficient usage within Site A (at least 1000 patients) to be used as targets for stability prediction (Supplemental eTable 1).

Covariate Definition

For each patient, available sociodemographic covariates include sex and race (one-hot categorical) as well as date of the visit and age of the patient (numerical). Additional patient covariates include all available coded billing data (i.e. ICD9 and ICD10 diagnoses, CPT lab tests and procedures), as well as the identity of all prescribed medications. From this initial set of 36,875 possible codes (or “codewords”), we selected 9,256 codewords which occurred in at least 50 patients in Site A. Thus, a count vector of size 9,256 represents a patient’s diagnostic and treatment history.

Baseline Methods for General and Drug-Specific Stability

The primary aim of prediction was to identify patients likely to exhibit general stability on antidepressants. That is, given the patient’s history up to an evaluation date, predict if the patient will be stable following index prescription of any antidepressant treatment. The secondary aim was to determine whether a subject would exhibit drug-specific stability.

One classifier was trained for the general stability outcome, as well as a separate drug-specific classifier for each of the 11 target antidepressants. We considered two standard probabilistic classifiers, logistic regression and extremely randomized trees, using the open-source implementations in Scikit Learn[24]. All classifiers were trained on Site A’s training set and had hyperparameters selected via grid search on Site A’s validation set to maximize area-under-the-ROC-curve (AUC). Final performance was evaluated on both Site A’s testing set and the independent cohort from Site B.

Supervised Topic Models for General and Drug-Specific Stability Prediction

A challenge in machine learning is maintaining interpretability while maximizing predictive performance. Even after applying the frequency threshold, an input space of 9,256 codewords limits interpretability and risks model overfitting. We therefore reduced this coded dataset into groups of co-occurring codes indicative of an underlying concept using probabilistic topic models[25]. For a diagram, see Supplemental eFigure 4.

We apply a recent technique for training topic models to perform supervised predictions called *prediction-constrained* topic modeling[21]. Most topic models simply summarize the most salient concepts in the data. For example, diseases such as diabetes, chronic kidney disease, and cancer are prevalent in health records and thus will always be ‘discovered’ as topics. However, it is not clear a priori whether these prominent conditions are relevant to predicting treatment response in

MDD - even given the importance of comorbidity, solely rediscovering comorbidity might exclude other features important for prediction. Prediction-constrained (PC) topic models address this issue, finding concepts useful for specific prediction tasks rather than simply summarizing prominent elements. We intend PC topic models to provide low-dimensional patient-specific covariates that yield comparable performance to classifiers that use high-dimensional codeword covariates. However, we hope our topic covariates provide more interpretable insights into how elements of the patient history factor into prediction. For further discussion of topic modeling applications to coded clinical data, see work by McCoy et al.[26,27].

Following our prior work[21], we applied prediction-constrained training to fit supervised Latent Dirichlet Allocation topic models (PC-sLDA) to Site A's training set. We selected 10 topics as representing the best trade-off between validation performance and model size. Experimental details for training and hyperparameter selection, as well as visualization links for trained models, are included in the Supplement. Open-source code is available: https://github.com/dtak/prediction-constrained-topic-models.

Methods for Evaluating Suitability of Models for Medication Prioritization

We further sought to assess how drug-specific models could be used to select medications to prioritize for each patient and compare this to clinical practice. Evaluating such prioritized medications requires certain assumptions, because for most patients we only observed outcomes with one or a few of the 11 possible medications. Given the top 3 suggested medications for a patient, we can assign one of three categories: "unassessable" (none of the 3 have known stability outcomes for that patient), "assessable and stable" (at least one of the top 3 has a positive outcome), and "assessable and non-stable" (none of the top 3 is stable and at least one is non-stable). We can then across a population compute "top-3 stability accuracy" indicating the fraction of assessable patients that would be stable. This evaluation represents a biased but potentially useful proxy for a possible future usage of drug-specific models (biased because models were not trained to prioritize among medications).

Methods for Evaluating Models as Tools for Forecasting Needed Medication Changes

Finally, we evaluated models of general stability by assessing how well they could forecast the number of medication changes that an individual will require before stability is achieved. For each model, we predict a probability score for each patient in Site A's test set, use this to stratify subjects into 4 quartiles, then report for each quartile the average number of medication initiations observed in practice before achieving stability.

Results

Our cohort contained 81630 adults (69% female; age range 18-80 with mean age 48.46) across both sites who met our inclusion criteria based on diagnosis and treatment duration (Supplemental eTable 1). After further filtering out 4,133 patients who lacked any code history before the first visit and thus cannot have personalized predictions as well as 273 subjects from Site B who had no outcomes for the 11 target antidepressants, we had 51,048 patients at Site A (67% female, mean age 48.50) which were further divided into training, validation, and testing sets. Similarly, we had 26,176 patients in Site B (74% female, mean age 48.96) for external evaluation of models. Sociodemographics are summarized in Supplemental eTable 2, with further descriptive statistics in Supplemental eFigure 5.

Stability Outcome Prevalence and Face Validity

For psychiatrist-treated patients (n=11,985 at Site A), we observed that 2,642 (22%) never reached stability, 5,274 (44%) reached stability on the index prescription, and 4,069 (34%) reached stability eventually. In contrast, for primary care patients (n=14208 at Site A), we observed that 14,208 (34%) were never stable, 19,867 (48%) were stable following index prescription, and 7,583 (18%) were eventually stable. Overall (n=53,643 at Site A), we observed that 16,850 (31%) of patients were never stable, 25,141 (47%) of patients reached stability on the index prescription, and 11,652 (22%) reached stability eventually.

Comparison of Feature Representations on Site A

Figure 1 compares general and drug-specific prediction models for two possible feature representations: high-dimensional codeword count vectors plus demographics (“dem+words”) and the low-dimensional “topics” covariates provided by our PC-sLDA model. General stability performance was best with “dem+words” features and an ensemble of 512 decision trees, achieving an average AUC of 0.661 (95% CI 0.648 - 0.672). When using a simpler logistic regression classifier, the high-dimensional “dem+words” features yield an average AUC of 0.628 (95% CI 0.614 - 0.639). The 10-covariate topic representation captures much of this discriminative capability even when using simple logistic regression, achieving average AUC 0.627 (95% 0.615 - 0.639). See Supplemental eFigure 6 and eTables 3, 4, 5, and 6 for thorough comparisons of all feature-classifier combinations on Site A.

Generalization of Stability Outcome Predictions to Site B

Next, we examined the transferability of models trained on data from Site A to separate patients from Site B (see Supplemental eTable 2 for sociodemographics). Distribution of stability outcomes for Site B was similar: across all 27,987 subjects 13,018 (47%) were stable on the index prescription, 5,492 (20%) were stable eventually, and 9,477 (34%) were never stable, compared to 47%, 22%, and 31% at Site A.

Figure 2 shows general stability prediction for both Site A and Site B, again comparing high-dimensional “dem+words” features to our 10-dimensional topic features. Models trained on Site A transfer to Site B with only modest decay in AUC for both feature representations. Using “dem+words” features, average AUC was 0.661 (0.648 - 0.672) for Site A and 0.663 (0.654 - 0.671) for Site B. Using our 10-dimensional topic features, average AUC was 0.627 (0.615 - 0.639) for Site A and 0.619 (0.610 -0.627) for Site B. As an alternative evaluation, Supplemental eFigure 7 plots positive predictive value versus negative predictive value for each model and site.

Model Interpretability Qualitative Evaluation

We sought to understand which features were important for stability prediction. Figure 3 visualizes representative topics learned by our proposed 10-topic model for general stability. All topics demonstrated sufficient coherence to enable a qualitative description annotated by a clinician (RHP). For example, while topics 5 and 7 both capture routine primary care visits, topic 5 reflects more terms related to psychiatric evaluation, suggesting more aggressive intervention or more severe illness. Topic 1 includes terms indicative of treatment resistance. Topics 2 and 4 capture gynecologic outpatient practice and menopause, respectively. Supplemental materials include hyperlinks to an online visualization tool to explore the important features of all trained models; Supplemental eFigure 8 shows important features for the “dem+words” classifiers.

Evaluation of Medication Prioritization and Comparison with Clinical Practice

Next, we evaluated the top-3 stability accuracy achieved by models used to prioritize antidepressants for a patient (Supplemental eTable 7). When always predicting the same 3 medications most commonly stable in Site’s A training set, we measured top-3 stability accuracy at 0.602 (95% CI: 0.591-0.612; 64.1% of the 12762 patients in Site A’s test set were assessable). For observed clinical practice (in which many treatments prescribe just one medication, but some prescribe more), the top-3 stability accuracy was 0.602 (95% CI: 0.593-0.611; 99.5% of 12762 patients assessable). This improved to 0.637 (95% CI: 0.628-0.646; 99.8% of 12762 patients assessable) if we allow prescriptions with less than 3 medications to be filled up to 3 total by selecting from the most commonly stable antidepressants. By comparison, our extremely randomized trees model using all demographic and diagnostic code features achieved a top-3 accuracy of 0.622 (95% CI: 0.610-0.634; 47.4% of 12762 patients assessable). Performance with our topic model was poorer: top-3 accuracy was 0.581 (95% CI: 0.566-0.594; 38.2% of 12762 patients assessable).

Evaluation of General Stability Prediction to Forecast Needed Medication Changes

Finally, we assigned all individuals in the test set to a stability risk quartile by their general stability probability score (Supplemental eTable 8). For our extremely randomized tree model using all demographics and codewords, those in the top quartile had a mean number of additional medication trials of 0.736 (95% CI 0.688-0.796) beyond the initial prescription at first visit to achieve stability. Those in the bottom quartile require an average of 1.754 medication trials (95% CI 1.681-1.843) beyond the initial prescription to achieve stability. By comparison, using our topic model features and LR classifier, the top quartile has a mean number of additional medication trials of 0.864 (95% CI: 0.816-0.918), while the bottom quartile has a mean of 1.722 (95% CI: 1.647-1.799).

Discussion

In this analysis of electronic health records from more than 81,000 individuals across two health systems, we identified machine learning models that predict achievement of treatment stability, a proxy for effectiveness, based solely on coded clinical data already available rather than incorporating research measures or questionnaires.

At first blush, the discrimination we report is modest, with AUC's in the 0.60 - 0.66 range. A key question to ask, however, is 'compared to what?' We were unable to identify any similar published studies in generalizable cohorts, so we cannot make a direct comparison to another method. While an AUC of 0.8 is often seen as a magic number, we and others have argued that this makes little sense, because the necessary discrimination depends critically on the context in which prediction will be applied[28,29].

We also found that, contrary to our hypothesis, development of treatment-specific predictors rather than general predictors did not meaningfully improve prediction. This may reflect the observation that much of antidepressant response may be considered placebo-like, or nonspecific. That is, while antidepressants consistently demonstrate superiority to placebo[1], placebo response is substantial such that nonspecific predictors may outperform drug-specific ones. This result is consistent with the lack of success of efforts to find treatment-specific pharmacogenomic predictors[30]. Our results do not preclude the existence of such medication-specific predictors but suggest that other strategies may be required to identify them.

We also present a framework for understanding the behavior of our drug-specific models if used to guide antidepressant selection, comparing performance to observed clinical practice and to a baseline in which all patients receive the most common antidepressants. It bears emphasis that this represents an instance of transfer learning: the models were not trained to recommend antidepressants per se, nor to mimic clinician performance. However, it illustrates a likely application of these models in practice to personalize treatment selection. We find that the difference between clinician performance and suggesting the one-size-fits-all most common medications is remarkably modest - on the order of 3%. In light of the known similarities in efficacy between standard treatments, essentially all of which are derived from a common set of assumptions about monoaminergic neurotransmission, this is not surprising. Despite enthusiasm about personalized medicine, the hypothesis that personalization improves outcome has rarely been rigorously tested. However, the observation that our best models already yield results similar to clinicians suggests that clinical performance may not be as out of reach as AUCs alone might indicate.

Our analysis also suggests that general stability prediction may still be useful for stratifying patients and understanding personalized chances of stability. We describe an approach to estimating the number of treatment trials that may be avoided, or saved, where models are applied. The top quartile of predicted stability requires about 1 less medication trial than the bottom quartile, which suggests that devoting more care resources - for example, more intensive care management or scalable evidence-based therapies - to those in the lower quartiles might be a worthy targeted investment.

Our results further indicate that, while topic modeling does not improve prediction over high-dimensional representations, it does yield readily-interpretable concepts relevant to prediction. EHR data is widely acknowledged to be noisy, with codes applied inconsistently even by individual clinicians; it is easy to learn predictors that are capturing a site effect or that simply serve as proxies for some other variable. (Conversely, the individual coded terms are inconsistent between linear and nonlinear models, and many are difficult to align with clinical practice, further illustrating the advantage in interpretability of topic-based models.) The approach we describe, which maps EHR dimensions into interpretable topics, allows stakeholders to easily inspect the learned topic features to understand what co-occurring codeword features in patient history drive predictions. This property is critical for researchers seeking to understand more complex models, but also ultimately for clinicians who may employ them: nominating treatments without understanding why they are favored is unlikely to be accepted by clinicians accustomed to their own strain of personalization[31]. The transferability of our results to a second health system suggests a further advantage, namely that topics may be more robust to overfitting than individual token-based approaches. In other words, if the goal is to build models that generalize across health systems, supervised topics may help to avoid the tendency of code-based models to fit site-specific use of individual procedure or diagnostic codes.

We emphasize multiple important limitations in interpreting these results. The outcome we examined, stability, is markedly different from standard outcomes in clinical trials like remission or 50% reduction in symptoms. The standard approach to using EHR data has attempted to impose a clinical trial-like structure and outcome measure – that is, to extract or impute measures of depression severity. Some studies have sought to emphasize a common primary care depression screening tool such as the PHQ-9[32], which characterizes symptom frequency, not severity, and was not designed to measure response. Others rely on text from narrative clinical notes[16]. However, these approaches minimize the strengths (availability of large scale, if imperfect, data that corresponds to real-world experience) while emphasizing the weaknesses (lack of precision in diagnosis and symptom measurement) in health records. Moreover, they perpetuate the myth that depression symptoms are purely episodic; in reality, such symptoms tend to wax and wane over time for many patients.

In contrast to these past efforts, we employed a simple metric to assess stability based on historical prescribing data, relying on the fact that clinician and patient behavior is likely to be reasonable: effective and well-tolerated treatments are continued, while ineffective or intolerable medications are discontinued. We are attempting to answer a simple question: “If I write a prescription today, how likely am I to continue writing it for the next 90 days?”

In sum, these results should be considered a starting point: incorporation of additional outcomes, as well as additional clinician- and patient-level predictors, should only improve the quality of prediction. Indeed, improving prediction of individual treatment response will almost certainly require data from multiple modalities. If such predictors are integrated with coded data to form topics, it should be possible not only to achieve greater discrimination while preserving portability, but also to understand the key features driving that discrimination in a way not possible with other machine learning strategies. Once such models emerge, prospective investigation will be needed to determine to what extent they meaningfully improve outcomes, if at all. The present study is a necessary first step towards such investigation.

Acknowledgements

This study was funded by Oracle Labs and the National Institute of Mental Health (1R01MH106577-01) (R56MH115187). We thank Victor Castro for preparing the deidentified EHR dataset and Partners Research Computing for computational resources. We thank the Center for Research on Computation and Society (CRCS) and Harvard Data Science Initiative (HDSI) for sponsoring Melanie F. Pradier.

Role of the Funding Source

No funding source contributed to any aspect of study design, data collection, data analysis, or data interpretation. The corresponding author (MCH) had full access to all the data in the study. All authors shared the final responsibility for the decision to submit for publication.

References

1. Cipriani A, Furukawa TA, Salanti G, 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. *Lancet*. 2018;391(10128):1357-1366. doi:10.1016/S0140-6736(17)32802-7

2. 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):449-459.

3. Kroenke K, West SL, Swindle R, et al. Similar effectiveness of paroxetine, fluoxetine, and sertralin in primary care: a randomized trial. *JAMA J Am Med Assoc*. 2001;286(23):2947-2955.

4. Rush AJ, Trivedi MH, Wisniewski SR, et al. Bupropion-SR, sertraline, or venlafaxine-XR after failure of SSRIs for depression. *N Engl J Med*. 2006;354(12):1231-1242. doi:10.1056/NEJMoa052963

5. Papakostas GI, Fava M. Predictors, moderators, and mediators (correlates) of treatment outcome in major depressive disorder. *Dialogues Clin Neurosci*. 2008;10(4):439-451.

6. Fava M, Uebelacker LA, Alpert JE, Nierenberg AA, Pava JA, Rosenbaum JF. Major depressive subtypes and treatment response. *Biol Psychiatry*. 1997;42(7):568–576.

7. Fava M, Rosenbaum JF, Pava JA, McCarthy MK, Steingard RJ, Bouffides E. Anger attacks in unipolar depression, Part 1: Clinical correlates and response to fluoxetine treatment. *Am J Psychiatry*. 1993;150(8):1158-1163. doi:10.1176/ajp.150.8.1158

8. Perlis RH, Uher R, Ostacher M, et al. Association between bipolar spectrum features and treatment outcomes in outpatients with major depressive disorder. *Arch Gen Psychiatry*. 2011;68(4):351-360. doi:10.1001/archgenpsychiatry.2010.179

9. Uher R, Perlis RH, Henigsberg N, et al. Depression symptom dimensions as predictors of antidepressant treatment outcome: replicable evidence for interest-activity symptoms. *Psychol Med*. 2012;42(5):967-980. doi:10.1017/S0033291711001905

10. Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*. 2016;3(3):243-250. doi:10.1016/S2215-0366(15)00471-X

11. Noma H, Furukawa TA, Maruo K, et al. Exploratory analyses of effect modifiers in the antidepressant treatment of major depression: Individual-participant data meta-analysis of 2803 participants in seven placebo-controlled randomized trials. *J Affect Disord*. 2019;250:419-424. doi:10.1016/j.jad.2019.03.031

12. Jha MK, South C, Trivedi J, Minhajuddin A, Rush AJ, Trivedi MH. Prediction of Acute-Phase Treatment Outcomes by Adding a Single-Item Measure of Activity Impairment to Symptom Measurement: Development and Validation of an Interactive Calculator from the STAR*D and CO-MED Trials. *Int J Neuropsychopharmacol*. 2019;22(5):339-348. doi:10.1093/ijnp/pyz011

13. Nierenberg AA, Feinstein AR. How to evaluate a diagnostic marker test. Lessons from the rise and fall of dexamethasone suppression test. *JAMA*. 1988;259(11):1699-1702.

14. Thase ME. Using biomarkers to predict treatment response in major depressive disorder: evidence from past and present studies. *Dialogues Clin Neurosci*. 2014;16(4):539-544.

15. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *Npj Digit Med*. 2018;1(1):18. doi:10.1038/s41746-018-0029-1

16. Perlis R, Iosifescu D, Castro V, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol Med*. 2012;42(01):41–50.

17. McCoy TH, Castro VM, Cagan A, Roberson AM, Kohane IS, Perlis RH. Sentiment Measured in Hospital Discharge Notes Is Associated with Readmission and Mortality Risk: An Electronic Health Record Study. Ramagopalan SV, ed. *PLOS ONE*. 2015;10(8):e0136341. doi:10.1371/journal.pone.0136341

18. Freitas AA. Comprehensible Classification Models – a position paper. *ACM SIGKDD Explor Newsl*. 2014;15(1):10.

19. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-215. doi:10.1038/s42256-019-0048-x

20. Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv170208608 Cs Stat*. February 2017. http://arxiv.org/abs/1702.08608.

21. Hughes MC, Hope G, Weiner L, et al. Semi-Supervised Prediction-Constrained Topic Models. *Proc 21st Int Conf Artif Intell Stat AISTATS*. 2018;PMLR 84:1067-1076.

22. Murphy SN, Mendis M, Hackett K, et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. *AMIA Annu Symp Proc*. 2007:548-552.

23. Raebel MA, Schmittdiel J, Karter AJ, Konieczny JL, Steiner JF. Standardizing Terminology and Definitions of Medication Adherence and Persistence in Research employing Electronic Databases. *Med Care*. 2013;51(8 0 3). doi:10.1097/MLR.0b013e31829b1d2a

24. Pedregosa F, Veroquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.

25. Blei DM. Probabilistic topic models. *Commun ACM*. 2012;55(4):77-84. ACM New York, NY, USA. doi:10.1145/2133806.2133826

26. McCoy TH, Castro VM, Snapper L, et al. Polygenic loading for major depression is associated with specific medical comorbidity. *Transl Psychiatry*. 2017;7(9):e1238. doi:10.1038/tp.2017.201

27. McCoy TH, Castro VM, Snapper LA, Hart KL, Perlis RH. Efficient genome-wide association in biobanks using topic modeling identifies multiple novel disease loci. *Mol Med*. 2017;23:285-294. doi:10.2119/molmed.2017.00100

28. Perlis RH. Translating biomarkers to clinical practice. *Mol Psychiatry*. 2011;16(11):1076-1087. doi:10.1038/mp.2011.63

29. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction.

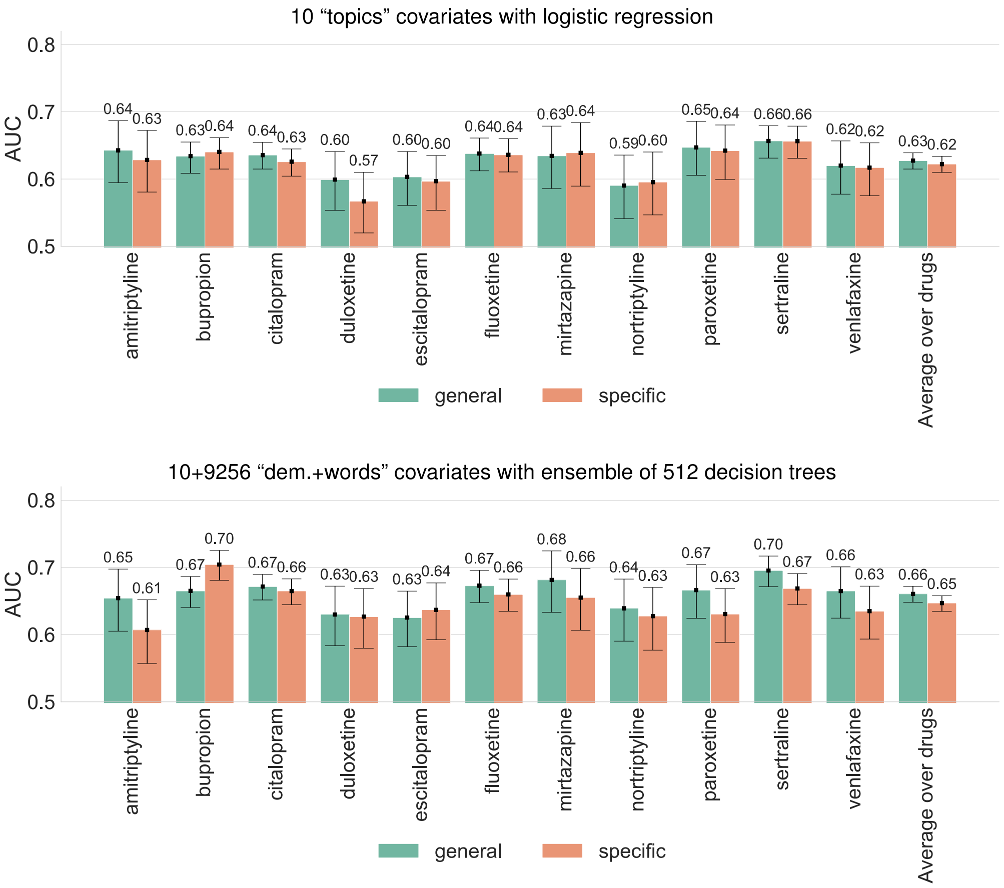
Circulation. 2007;115(7):928-935. doi:10.1161/CIRCULATIONAHA.106.672402

30. GENDEP Investigators, MARS Investigators, STAR*D Investigators. Common genetic variation and antidepressant efficacy in major depressive disorder: a meta-analysis of three genome-wide pharmacogenetic studies. *Am J Psychiatry*. 2013;170(2):207-217. doi:10.1176/appi.ajp.2012.12020237

31. Perlis RH. Abandoning personalization to get to precision in the pharmacotherapy of depression. *World Psychiatry*. 2016;15(3):228-235. doi:10.1002/wps.20345

32. Kroenke K, Spitzer RL. The PHQ-9: A new depression diagnostic and severity measure. *Psychiatr Ann*. 2002;32(9):509-515. doi:10.3928/0048-5713-20020901-06

Figure 1: Comparison of General and Drug-specific Stability Prediction for Proposed and Baseline Covariates


10 “topics” covariates with logistic regression

10+9256 “dem.+words” covariates with ensemble of 512 decision trees

Figure 1: Comparison of discriminative ability, as measured by area under the ROC curve (AUC), for general and drug-specific prediction models. Top panel compares models that utilize our proposed 10-dimensional “topics” covariates with a logistic regression predictor. Bottom panel compares models that utilize the baseline high-dimensional “dem+words” covariates with an ensemble of 512 extremely randomized decision trees. For each of the 11 target antidepressants, we obtain an AUC score for a given model by considering predictions from that model on the subset of the Site A test set that includes all known outcomes related to that drug (ignoring data from patients who were never given that drug). To indicate uncertainty in reported AUC values, we repeat our evaluation across 5000 bootstrap samples of each test set and report error bars indicating 95% confidence intervals for the AUC across these bootstrap samples (i.e. the 2.5 and 97.5 percentiles).

Figure 2: Assessment of Generalization from Site A to Site B for both General and Drug-specific Stability Prediction

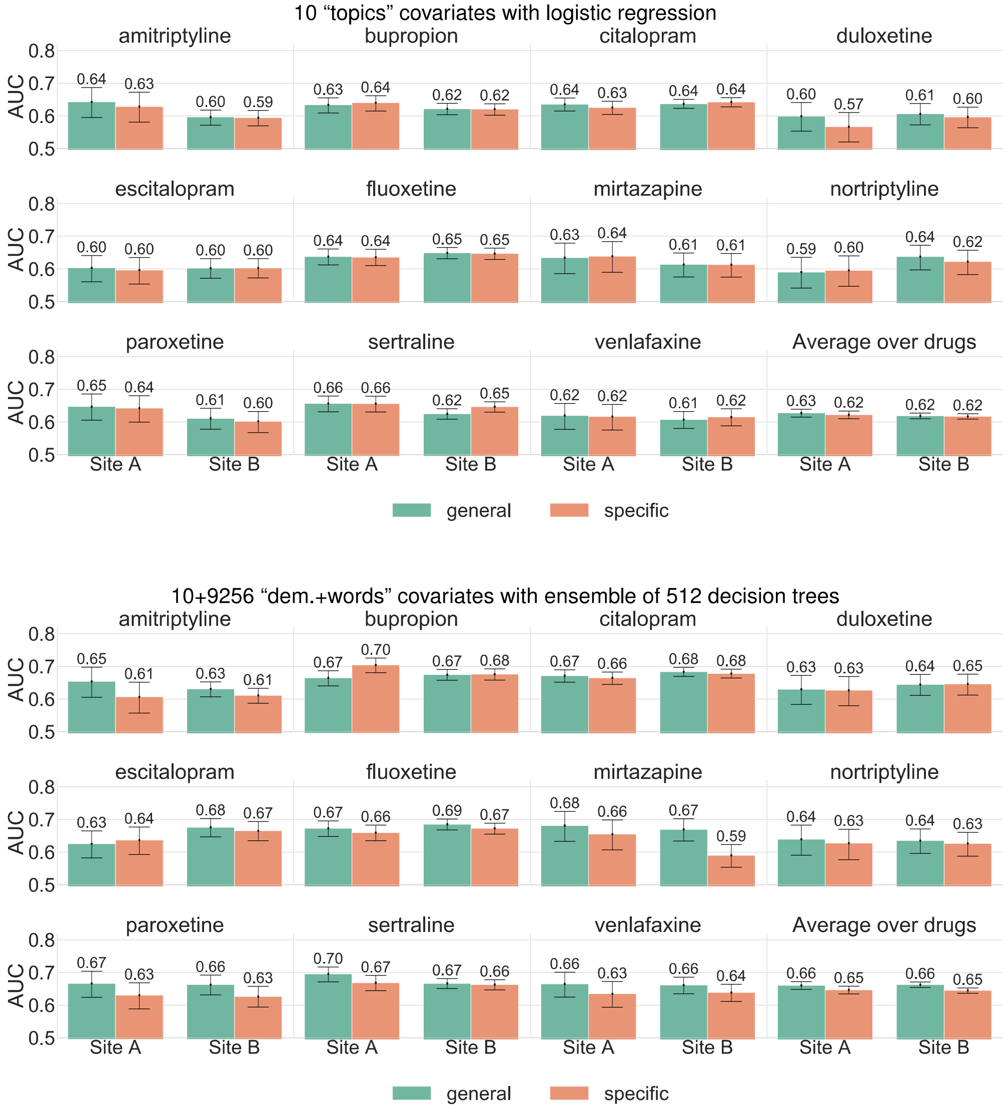


Figure 2: Side-by-side comparison of discriminative ability on the Site A and Site B testing sets, as measured by area under the ROC curve (AUC), for general and drug-specific prediction models. Top panel compares models that utilize our proposed 10-dimensional "topics" covariates with a logistic regression predictor. Bottom panel compares models that utilize the baseline high-dimensional "dem.+words" covariates with an ensemble of 512 extremely randomized decision trees. For each of the 11 target antidepressants, we obtain an AUC score for a given model by considering predictions from that model on the subset of the Site A test set that includes all known outcomes related to that drug (ignoring data from patients who were never given that drug). To indicate uncertainty in reported AUC values, we repeat our evaluation across 5000 bootstrap samples of each test set and report error bars indicating 95% confidence intervals for the AUC across these bootstrap samples (i.e. the 2.5 and 97.5 percentiles).

Figure 3: Visualization of Supervised Topic Model for General Stability

Topic #5: "Primary Care with Some Psychiatry"
LR Coef. = - 1.0 (Negative correlation w. stability)

Proba.	*Type*	*ID*	*Word*
0.033	CPT	99213	Office visit > 15 min.
0.024	CPT	99214	Office visit > 25 min.
0.016	CPT	99211	Office visit > 05 min.
0.015	CPT	08527	Complete blood count tests
0.010	CPT	82565	Creatinine blood test
0.010	ICD	78900	Abdominal pain
0.010	CPT	85025	Complete blood count tests
0.010	CPT	71020	Radiologic exam of chest
0.009	CPT	84520	Urea nitrogen lab test
0.009	Rx	42347	Bupropion prescription
0.009	ICD	311	Depressive disorder

Topic #9: "Back / Joint Pain"
LR Coef. = -0.1 (Negative correlation w. stability)

Proba.	*Type*	*ID*	*Word*
0.098	CPT	97110	Physical therapy
0.052	ICD	7245	Back ache
0.037	CPT	97140	Manual therapy
0.033	Rx	7804	Oxycodone
0.032	ICD	7242	Lumbago
0.019	Rx	214182	Acetaminophen/Hydrocod.
0.018	ICD	7231	Cervicalgia
0.015	ICD	71941	Shoulder pain
0.012	Rx	25480	Gabapentin
0.012	ICD	71947	Ankle/foot pain
0.012	ICD	71596	Osteoarthrosis of lower leg

#2 "Primary Care for Younger Women"
LR Coef. = + 0.0 (Neural correlation w. stability)

Proba.	*Type*	*ID*	*Word*
0.034	CPT	87591	Test for gonorrhoeae
0.033	CPT	87491	Test for chlamydia
0.020	CPT	87070	Bacterial culture
0.017	CPT	81025	Urine pregnancy test
0.017	CPT	84702	hCG test
0.016	CPT	87086	Bacterial culture fr. urine
0.015	CPT	v762	Cervix screening
0.015	ICD	462	Acute pharyngitis
0.015	CPT	V222	Incidental pregnancy
0.015	CPT	76856	Pelvic ultrasound
0.015	ICD	6259	Female genital concern

Topic #4: "Primary Care for Older Women"
LR Coef. = + 0.1 (Neutral correlation w. stability)

Proba.	*Type*	*ID*	*Word*
0.048	CPT	v761	Mammogram
0.038	Rx	10582	Levothyroxine
0.032	ICD	2449	Hypothyroidism
0.025	ICD	2724	Hyperlipidemia
0.021	CPT	76092	Mammogram
0.020	CPT	v762	Cervix screening
0.019	ICD	6272	Menopause
0.017	ICD	73300	Osteoporosis
0.013	ICD	v103	Breast cancer history
0.012	ICD	6961	Psoriasis
0.012	ICD	78079	Malaise and fatigue

Topic #1: "Treatment Resistant MDD"
LR Coef. = + 0.3 (Positive corr. with stability)

Proba.	*Type*	*ID*	*Word*
0.138	ICD	29630	Major depressive disorder
0.125	CPT	90806	Psychotherapy
0.098	ICD	90862	Rx review psychotherapy
0.034	ICD	30000	Anxiety
0.032	Rx	2598	Clonazepam
0.023	ICD	29650	Bipolar disorder
0.023	ICD	30490	Drug dependency
0.022	CPT	90870	Electroconvulsive therapy
0.018	ICD	30981	PTSD
0.018	ICD	2967	Bipolar disorder
0.018	CPT	90807	Psychotherapy

Topic #7: "Primary Care"
LR Coef. = + 0.7 (Positive corr. with stability)

Proba.	*Type*	*ID*	*Word*
0.048	CPT	99213	Office visit > 15 min.
0.037	CPT	99214	Office visit > 25 min.
0.029	CPT	99211	Office visit > 5 min.
0.021	CPT	36415	Blood drawn for lab test
0.016	CPT	85027	Complete blood count test
0.016	CPT	v700	Routine exam
0.016	Rx	7646	Omeprazole
0.013	CPT	80061	Lipid panel
0.011	CPT	90658	Flu vaccine
0.011	CPT	99215	Office visit > 40 min.
0.011	CPT	80053	Metabolomic tests

Figure 3: Display of 6 learned topics from our proposed PC-sLDA supervised topic model trained to predict general stability. We selected these 6 topics as representative of the 10 total topics learned by the model (top row shows the two most negatively correlated with stability as indicated by each topic's learned logistic regression coefficient; middle row shows the two most neural; and bottom row shows the two most positively correlated with stability). Each topic is defined by a learned distribution over 9,256 possible diagnostic (ICD), procedural (CPT), and medication-related (Rx) codewords. Codewords with high probability in the same topic are likely to co-occur together in a patient's record explained by that topic. For each topic, we show its top 10 most probable codes. Each topic is labeled with a clinician annotated title (provided post-hoc by RHP), and the topic's index order within the original model. We also show the learned logistic regression coefficient (rounded to the nearest 0.1) for the task of predicting general stability. Large positive coefficients suggest that a patient whose history uses more of that topic will be more stable. For an online visualization of all 10 topics, see our visualization webpage.