Evaluation of machine learning methods for prognostic and adverse event risk in Major Depressive Disorder

By

Barrett Jones

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Informatics

May 10, 2024

Nashville, Tennessee

Approved:

Colin G. Walsh, MD, MA

Laurie Novak, PhD

Andrew J. Spieker, PhD

Warren Taylor, MD, MHS

Adam Wright, PhD

Acknowledgements

This work was supported by the National Library of Medicine training grant number T15 LM007450-17 and Wellcome Leap Multi-Channel Psych.

Table of Contents

List of Tables

List of Figures

Introduction

Major Depressive Disorder (MDD) is a prevalent illness that impacts many worldwide throughout their lifetimes. It has been cited as the third leading cause of disability world-wide,[1] and suicide is a top 10 leading cause of death among those aged 10-64.[2] The diagnostic criteria for depression takes into account symptoms such as mood, diminished pleasure, weight loss, fatigue, cognitive impairment, and suicidal ideation.[3] A depressive episode can be brought on by life stressors such as loss of a loved one, or divorce.[4,5] The biologic mechanism causing MDD is not well understood, function among neurotransmitters has been the suspected cause due to therapeutic response to antidepressant medications that alter function of these neurotransmitters.[1] An array of treatments have displayed efficacy in MDD, including psychotherapy, antidepressant medications, partial hospitalization, and brain stimulation therapies, although many patients have incomplete or no response. Antidepressants have shown efficacy in moderate to severe MDD, however many considerations go into making an effective choice in antidepressant therapy as several classes of drugs are available, with each class containing multiple medication options, and patients may experience side effects.[6] Additionally, MDD is heterogenous and treatment decisions are complicated by comorbid conditions such as anxiety, psychotic symptoms, substance abuse, and borderline personality disorder.[3]

The responsibility for a significant amount of depression care falls to non-mental health care providers.[7] Effective prognostic risk models can aid providers in treatment decisions.[8] The use of electronic health record (EHR) data for clinical risk modeling has become increasingly prevalent in recent years.[9] EHR data are high dimensional and complex, posing significant challenges. Some of the complexity inherent in EHR data includes temporality, noise, and sparsity, which can negatively impact predictive performance.[9–11] Embedding models may lead to performance gains in predictive models.[12] In Chapter 1, we evaluate the clinical usefulness of autoencoders in clinical risk prediction. Autoencoders are an unsupervised learning method that can represent feature dependencies in latent embeddings. In addition, autoencoders can denoise, reduce dimensionality, and reduce sparsity. We find that pretraining autoencoders and fine-tuning improves predictive performance relative to embedding models, neural networks without pretraining and aggregation-based feature representation methods. Improvement in prediction resulting from pretraining has potential for increased clinical impact of MDD risk models.[8] Additionally, our finding

that pretrained models outperform embedding models suggests that important information for prediction contained in the weights may not be passed to the embeddings.

Switching of antidepressants is commonly done in practice and adverse event risk is an important consideration when a new antidepressant is prescribed.[13,14] Electronic health record data capture routine care and are a potentially effective source for post-market safety surveillance. The observational nature of these data can result in biased analyses. In Chapter 2, we attempted to better understand and correct for bias using the empirical calibration with control outcomes method from Schuemie et al.[15] Empirical calibration models the relationship between study covariates and control outcomes for which the true treatment effect is likely known and fits a systematic bias model that both measures bias and calibrates population level treatment effects. We compared two common methods for treatment effect estimation, propensity score weighting and matching and found a protective bias in treatment effect estimates from propensity score weighting, which may be due to selection, healthy user, or some other form of bias. Matching results showed little evidence of bias, potentially due to this population selection being guided by clinical expertise.

Heterogeneity of antidepressant adverse events in MDD subpopulations is an important consideration in treatment decisions. Patients with higher comorbid burden have been shown to be less likely to be prescribed antidepressants.[16,17] Also, concerns about antidepressant efficacy and safety in MDD subpopulations have led to multiple clinical trials in various subpopulations.[14,18,19] Statistical analysis of heterogeneity can be tenuous and lead to spurious findings when many subpopulations are analyzed. Heterogeneous Treatment Effect (HTE) models reframe the problem to focus on the detection of heterogeneity, rather than conducting a stepwise subgroup analysis. In Chapter 3 we studied recent advances in machine learning allow for flexible modeling of HTE. We evaluated HTE modeling techniques under varying data generating processes with semi-synthetic outcomes—synthetically generated outcomes that use real data as a baseline. We also evaluated HTE models on adverse event outcomes. Analysis of semi-synthetic and real-world adverse event outcomes allowed us to first gain insight into performance of the HTE models under varying data generating processes and use this to inform interpretation of results in the adverse event outcomes. We observed variance in HTE model performance across data generating processes and outcomes and identified tuning strategies as a key area of research for advancement of HTE models.

We aim to advance capabilities to model prognostic and intervention risk in MDD by study of statistical and machine learning methods on observational EHR data. To accomplish this, we evaluated autoencoder ability to improve predictive performance through feature engineering and pretraining across multiple prediction tasks in MDD. We also employed methods for empirical calibration of treatment effects to estimate adverse event risk in antidepressant prescribing. HTE model performance was assessed under varying data generating processes with semi-synthetic outcomes, which then informed interpretation of real adverse event modeling results. Finally, we tested HTE model ability to detect adverse event heterogeneity.

Chapter 1 Sequential autoencoders for feature engineering and pretraining in Major Depressive Disorder risk prediction

Background and Significance

The use of electronic health record (EHR) data for clinical risk modeling has become increasingly prevalent in recent years.[9] However, EHR data are high dimensional and complex, posing significant challenges for effective analysis. Some of the complexity inherent in EHR data includes temporality, noise, and sparsity, which can negatively impact predictive performance.[9–11] To address these challenges, autoencoder models have emerged as a promising approach for generating simplified representations that reduce dimensionality, denoise, and account for temporality.[20–22] Moreover, pretrained weights can reduce training time and increase predictive performance.[23–25] Previous studies have shown that autoencoders and other pretrained encoding models can achieve state-of-the-art prediction accuracy in diagnostic tasks and may learn complex disease relationships.[9,26]

These approaches may be particularly important for common psychiatric disorders, including Major Depressive Disorder (MDD). Machine learning applications have been widely used for prognostic prediction to support clinicians in the identification of individuals with MDD at elevated risk for suicidal behavior.[27–33] Creating models with clinical benefit in this syndromal phenotype is particularly difficult. In a recent meta-analysis[29] a majority of risk models considered had a precision less than 1%, resulting in concerns about the clinical usefulness of suicidality risk models and a negative relationship between model performance and study quality may exist.[31] It has been shown that for cost effectiveness suicide-attempt models should exceed a precision of 0.8%.[8] Low predictive performance in these studies can in part be explained by class imbalance in training datasets, and lack of clear evidence for suicidality risk factors.[34] These studies present opportunity for innovative machine learning techniques to improve predictive performance and clinical benefit of risk models in the MDD population.

It is common practice for researchers working with EHR data to generate aggregate features for prediction of outcomes relevant to MDD. Due to the sparsity of outcomes and features in the patient population, autoencoders have potential to improve predictive performance. Tran et al.[12] show Restricted Boltzman Machine (RBM) encodings improved prediction in patients under suicide risk assessment. A recent review of deep learning techniques for automated feature representation identified 49 recent publications in which automated feature

representation was applied to a range of prediction tasks.[9] Autoencoder pretraining has been shown to improve predictive performance in biomedical prediction tasks.[24,25] Autoencoder feature engineering has also been applied to patient subtyping,[9] treatment trajectory characterization,[22] and causal inference.[35–37]

Autoencoders are composed of encoder and decoder sub-models that have the capability to represent complex feature dependencies.[20] The encoder model, f_θ, maps the inputs, x, to a lower-dimensional latent space representation, $z = f_\theta(x)$. The decoder model, $g_{\theta'}$, maps this latent vector back to the original input space to reconstruct the input $x' = g_{\theta'}(z)$. The model is fit by minimizing the error between the original input and reconstructed inputs. Autoencoders can flexibly accommodate various encoder and decoder model architectures. To account for the temporal nature of EHR data, this work focuses on the use of sequential neural networks.

Objective

The objective of this study was to evaluate autoencoder ability to improve predictive performance through feature engineering and pretraining across multiple prediction tasks in MDD. We evaluated autoencoder model ability to capture temporal disease relationships that may not be identified by aggregate features and whether that information is retained in the model encodings, z, or the pretrained weight values f_θ. The predictive performance of autoencoder models of multiple structures were investigated in an array of clinical outcomes, including unplanned admissions, emergency department (ED) visits, high utilization, and self-harm/suicide attempt. The included health utilization outcomes may ease the challenges of suicidality risk prediction and maintain clinical relevance—through association with MDD severity.[38–40] To evaluate autoencoders as a feature engineering technique, encodings are input to a random forest model for prognostic prediction, as random forests have shown strong prediction performance in the MDD population in prior studies.[28,33] To evaluate autoencoder pretraining, encoder weights are extracted from the autoencoder models and used to initialize neural network prediction models. Predictive performance for autoencoder feature engineering were compared to benchmarks of a random forest trained on aggregate features and an RBM as in Tran et al.[12] Pretraining predictive performance is compared between LSTM and Attention neural network models of the same structure, but without pretraining, as well as the best performing feature engineering technique.

Methods

Data description

This study examined data from the Vanderbilt University Medical Center (VUMC) Research Derivative.[41] VUMC is located in the United States mid-south, Nashville, Tennessee. The Research Derivative includes data from multiple clinical systems that is structured for research purposes. Patients were included in the study having an MDD indication between 1/1/2013-12/31/2018. We defined MDD indication as a depression-related International Classification of Diseases (ICD) diagnosis code, including MDD, dysthymic disorder, and depressive disorder not elsewhere classified,[42,43] an antidepressant prescription, or problem list mention of depression. The ICD codes are in appendix table A.1 and list of antidepressants table A.2. We additionally required that patients had a depression-related ICD code during the time period of analysis, are 18-90 years old at indication, had two visits six months apart at VUMC prior to indication, and were not diagnosed with bipolar disorder or schizophrenia. Study data were extracted from the research derivative using IBM Netezza SQL and preprocessing was done in Python (version 3.8).

For patients meeting entry criteria we extracted 3 years of data after the initial MDD indication and formatted it into a quarterly time series with feature indicators. Feature categories included diagnoses, interventions, and outcomes. International Classification of Diseases (ICD) diagnosis codes[42] were extracted and grouped using Agency for Healthcare Research and Quality Clinical Classification Software (CCS)—a grouping of ICD codes intended to be clinically meaningful.[44]

Study interventions were identified and defined with clinical expert guidance (WDT) and were extracted from orders and notes data. Interventions included prescribing one or multiple antidepressants, antidepressant dose change, psychotherapy referral, partial hospitalization referral, and electroconvulsive therapy (ECT) referral. Interventions that were not related to medications were supplemented by notes data. Regular expressions were developed to identify mentions of a referral or consult for psychotherapy, ECT consultation, and partial hospitalizations in both the notes and orders data. Antidepressant data were extracted by Anatomical Therapeutic Chemical (ATC) code[45] according to a list of antidepressants identified by collaborating psychiatrists. The list of antidepressants has been included in Table A.2 in the appendix with ATC code.

Outcomes included self-harm/suicide attempt, unplanned admission, ED visit and high utilization. Unplanned admissions were defined as patients admitted to the hospital excluding any admissions that may be considered part

of planned treatment according to the Center for Medicare and Medicaid Services Unplanned Readmissions Algorithm.[46] High utilization was defined as any patient with two or more inpatient or emergency room visits with an MDD related ICD code during a quarter. The self-harm/suicide attempt outcome is ICD code based, where ICD codes were mapped to the self-harm/suicide attempt CCS code.

Autoencoder architectures

Autoencoders are composed of two sub-models—an encoder and decoder. The encoder takes the input data and outputs a latent representation. The latent representation is input to the decoder model from which it reconstructs the input data. We test neural network architectures that account for the sequential nature of time series data. LSTMs are a form of Recurrent Neural Network (RNN) that stores information over extended sequences and employs a gating method to address the exploding gradient problem found in some RNN applications.[47] Attention based architectures learn long term dependencies and have been shown to outperform LSTMs on natural language processing tasks.[48] In contrast to RNNs that keep a state representation that is updated at each position of the sequence, the attention mechanism identifies valuable past information given the current state.

The attention model architecture was adapted from Vaswani et al. to work with time series features.[48] The Keras (version 2.12.0) python library was used to construct the autoencoder models.[49] The Attention encoder block is composed of a multihead attention layer followed by two one dimensional CNN layers. The decoder block is composed of two multihead attention layers followed by a one dimensional CNN layer and a time distributed feed forward layer. The multihead attention layer computes multiple self-attention layers in parallel to attend to different parts of the input sequence simultaneously. The CNN layer learns a convolution kernel across the time series vector and is able to capture local patterns in the sequence.[50] The time distributed feedforward layer applies a fully connected layer to each timestep in the sequence. This layer has a sigmoid activation to estimate the probability of the original inputs. The LSTM encoder includes two LSTM layers in both the encoder and decoder blocks and time distributed feed forward layer in the decoder block. Figure 1 displays the structures of the attention and LSTM encoder and decoder blocks. Further details on layer parameters are included in the appendix Table A.3.

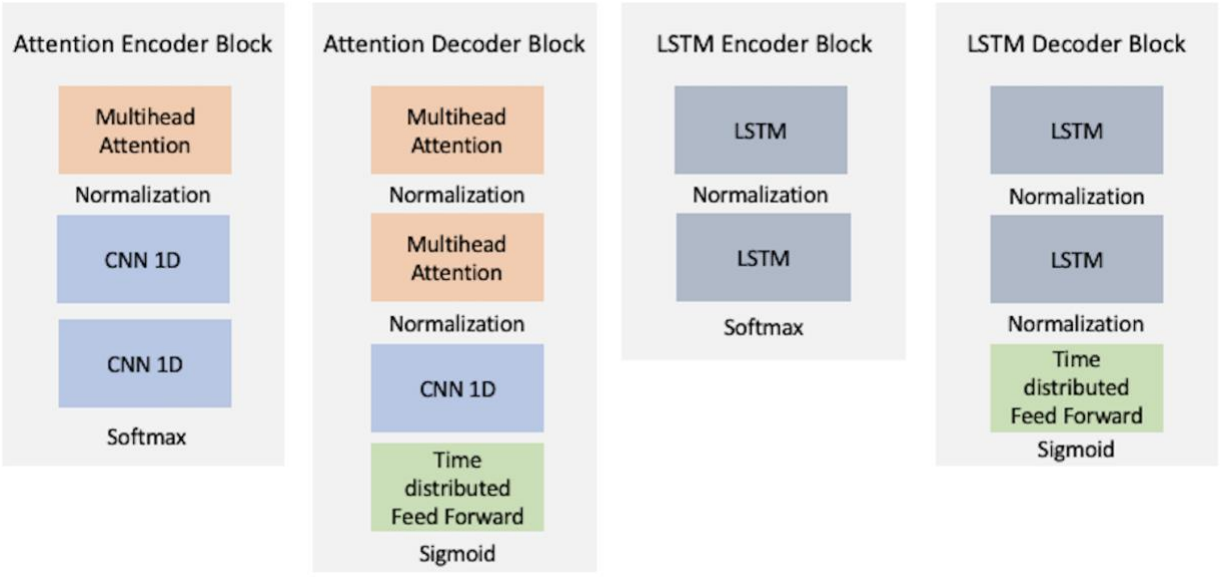


Figure 1-1 Attention and LSTM model encoder and decoder blocks.

Attention encoder block has a single multihead attention layer followed by two single dimension convolutional neural network (CNN) layers. The attention decoder has two multihead attention layers followed by a CNN and time distributed feedforward layer. The sigmoid activation on the final layer outputs probability estimates of the input indicators. The LSTM encoder block is composed of two LSTM layers and the decoder two LSTM layers followed by a time distributed feed forward layer that estimates the probabilities of the inputs.

Benchmark models

As a benchmark feature representation, the time series were aggregated by yearly rolling feature counts at each timepoint. A random forest model was fit with grid search cross validation of hyperparameters for each study outcome. RBM's have been shown to achieve state-of-the-art predictive performance in the MDD phenotype.[12] We evaluated an augmented version of the model of Tran et al. optimized to the prediction tasks of this study. Time series data were aggregated indicator features corresponding to (0–90), (90–180), (180-360) and (360–720) day intervals. An RBM with elasticnet regression pipeline was fit with concurrent hyperparameter tuning to optimize AUPRC. RBM tuned parameters included number of components and learning rate. LSTM and Attention neural networks were composed of the LSTM and Attention encoding blocks (see figure 1), without pretraining, followed by a time distributed dense layer. An LSTM and Attention model was fit for each outcome.

Model training and evaluation

Patients meeting entry criteria were split, two-thirds into a training and one-third into a test set. The quality of each representation is evaluated in each study outcome. Autoencoder models were fit on the training data and at each time point we fit a random forest model with grid search hyper parameter tuning with autoencoder latent vectors as input. Weights were extracted from the encoders of both the LSTM and Attention autoencoders. A time distributed dense layer was appended to the encoding layers and a predictive model was trained on each of the study outcomes. Area under the precision recall curve (AUPRC) was recorded for each predictive model on the test set. We recorded AUPRC at each time point on the test set, and trends and variation were evaluated.

Results

Of the 27,319 patients meeting entry criteria 17,621 (64.5%) were female. Most patients are non-hispanic/latinx whites (n=22,478, 82.3%). Black patients account for 11.0% of the population, with 1,027 (3.8%) of patients being classified as other, this includes patients with unreported or multiple reported races. The most common intervention in the study is prescribing an antidepressant (n=19,414, 71.1%), of those 42% are prescribed multiple. Table 1 contains details on the patient population, count data are aggregated across the entire study period.

Table 1-1 Patient descriptive statistics aggregated across the study period

		N	%
Gender			
	Female	17,621	64.5%
Race/Ethnicity			
	American Indian or Alaska Native	52	0.2%
	Asian	379	1.4%
	Black	2,996	11.0%
	White-Hispanic/Latinx	387	1.4%
	White-not Hispanic/Latinx	22,478	82.3%
	Other	1,027	3.8%
Age		Mean	SD
	Age	48.1	18.1
Outcomes		N	%
	Unplanned Admission	11,172	40.9%
	ED Visit	5,278	19.3%
	High Utilization	968	3.5%
	Self-harm/Suicide Attempt	2,032	7.4%
Interventions			

Antidepressant Prescription	19,414	71.1%
Multiple Antidepressant Prescriptions	8,300	30.4%
Dose Increase	1,838	6.7%
Dose Decrease	922	3.4%
Psychotherapy Referral	7,592	27.8%
ECT Referral/Consult	101	0.4%
Partial Hospitalization Referral/Consult	213	0.8%
Diagnoses	Mean	SD
CCS code count (unique)	14.3	10.3

Through the entire study period 11,172 (40.9%) patients have an unplanned admission, by quarter this outcome ranges in proportion from 5.1% to 8.7%. ED visit proportions by quarter range from 1.5% to 3.6% Each of the study outcomes had a decreasing trend across the study period (ordinary least square p-value < 0.001). High utilization and self-harm/suicide attempt are relatively infrequent, ranging from 0.20%-0.54% and 0.13%-0.31% respectively. Figure 2 shows outcome proportion trends by quarter.

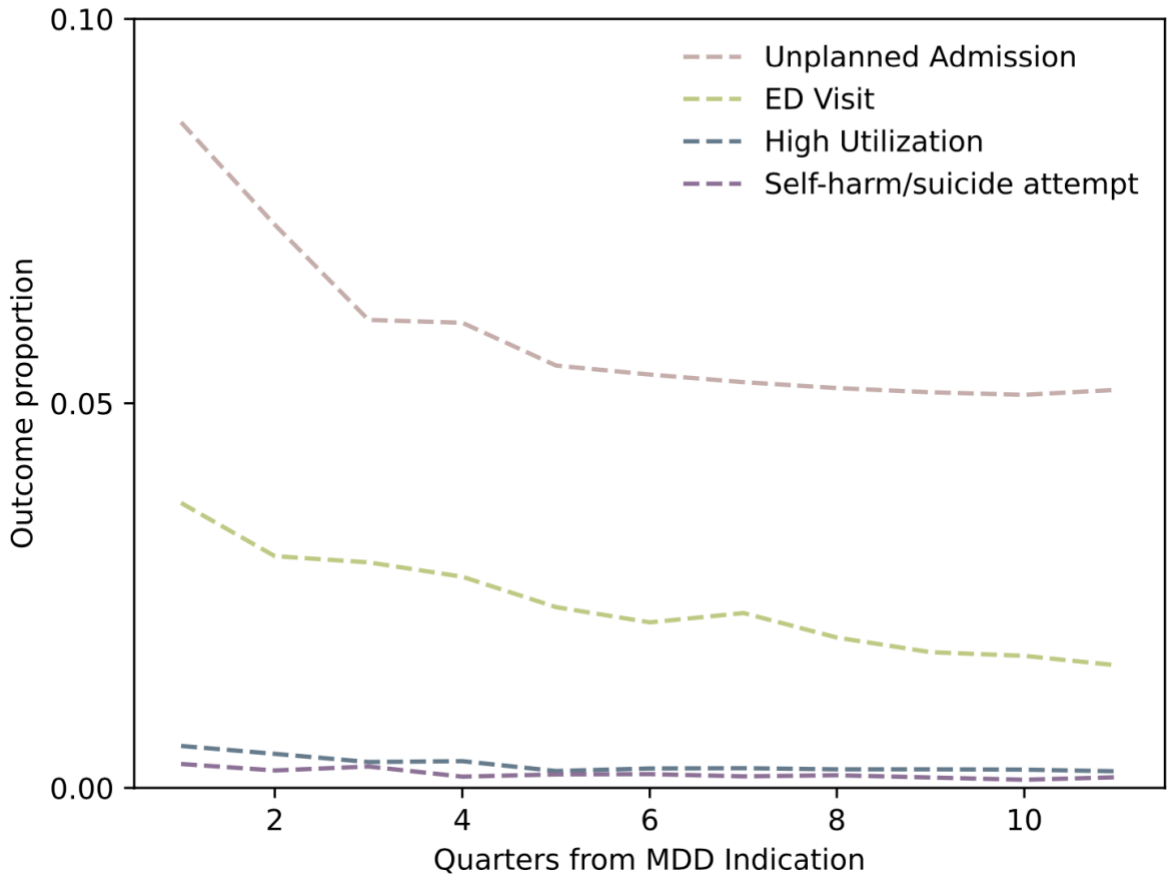


Figure 1-2 Outcome frequency trend by quarter.

The proportion of patients with the observed outcomes is plotted at each timepoint in the study.

The training set included 18,213 patients. Autoencoders were trained with a 10% validation set and binary cross entropy loss. Final validation error was 0.012 for the Attention autoencoder with 997,314 trainable parameters. Validation loss for the LSTM autoencoder was 0.006 with 717,066 trainable parameters.

The predictive performance of each autoencoder as a feature engineering method was evaluated temporally by fitting a random forest on model encodings with grid search cross validation parameter tuning. Test AUPRC scores were calculated at each time point and are reported in Figure 3. RBM had the highest average AUPRC across each prediction task, except for High Utilization, where the LSTM and Attention feature engineering had higher AUPRC (RBM 0.050, 95% CI 0.023-0.077, LSTM 0.059, 95% CI 0.031-0.087, Attention 0.062, 95% CI 0.038-0.085). RBM

had significantly better AUPRC for the ED visit outcome (0.16, 95% CI 0.14-0.17), where next best was aggregate feature engineering (0.11, 95% CI 0.10-0.12). Relatively high variance in AUPRC was observed in the self-harm/suicide attempt outcome—RBM a ranges from 0.0018-0.020. LSTM compared to Attention based autoencoder feature engineering was similar across outcomes. Attention had higher average AUPRC in the unplanned admission, high utilization and self-harm/suicide attempt outcomes.

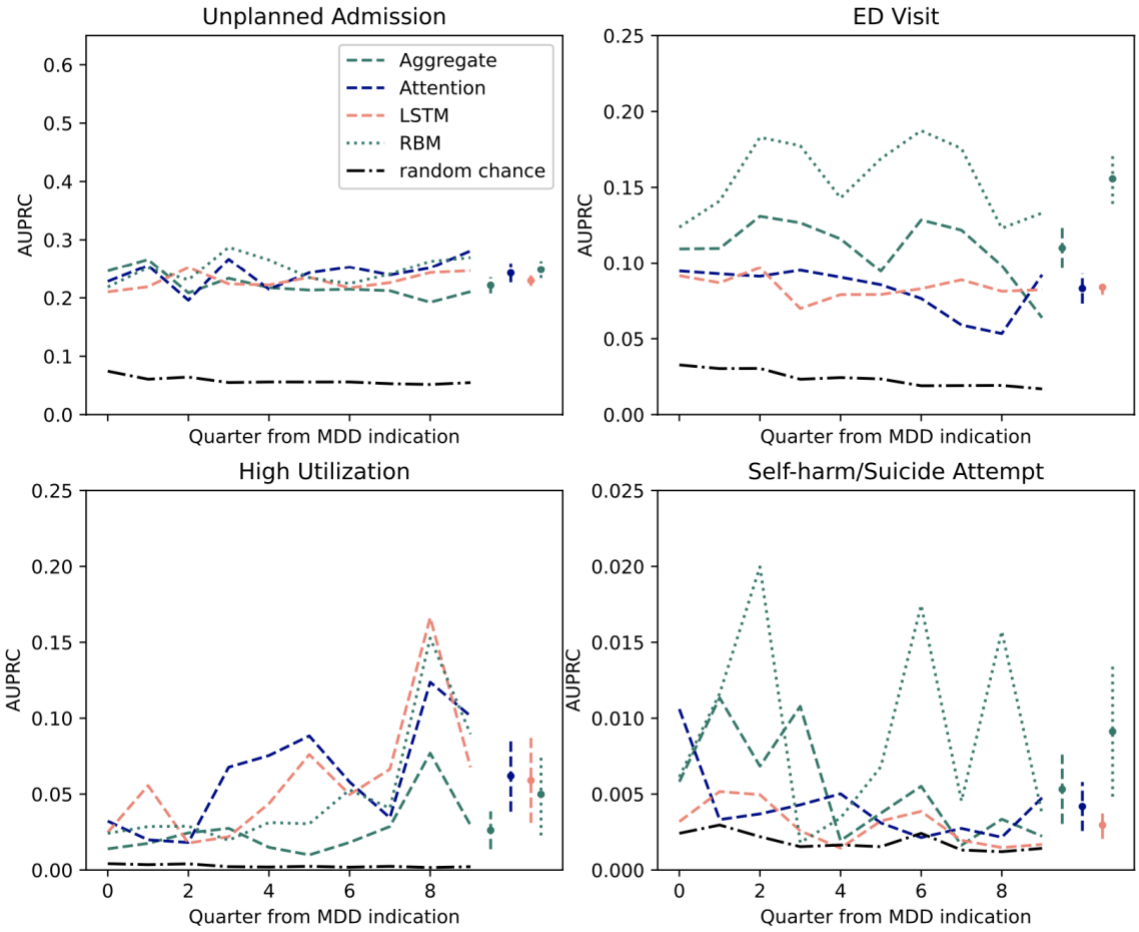


Figure 1-3 Temporal validation of feature engineering methods.

Each line displays the AUPRC trend for the corresponding feature engineering method. Plots are included for temporal validation of feature engineered training data at each quarter in the study. The black dot-dash line represents the performance of a random chance estimator. 95% confidence intervals are shown for the average AUPRC across the study time period.

The LSTM model with pretraining had the highest average AUPRC in three of four outcomes. The exception is the ED visit performance where RBM has the highest AUPRC (RBM=0.16, 95% CI 0.14-0.17, LSTM pretrained=0.14, 95% CI 0.12-0.15). Pretraining resulted in a increase in performance over LSTM without pretraining in each outcome. LSTM with pretraining had highest average AUPRC in the self-harm/suicide attempt outcome, but due to variation over time, the result is not a significant improvement over benchmark. Pretrained attention models had comparable performance relative to attention without pretraining. Attention without pretraining had higher average AUPRC in ED visit, and self-harm/suicide attempt. The self-harm/suicide attempt LSTM pretrained model had a precision of 1.45%, recall 30.08%, and specificity 95.31% for the top 5% of risk predictions in patients with observed features during the prediction quarter.[51]

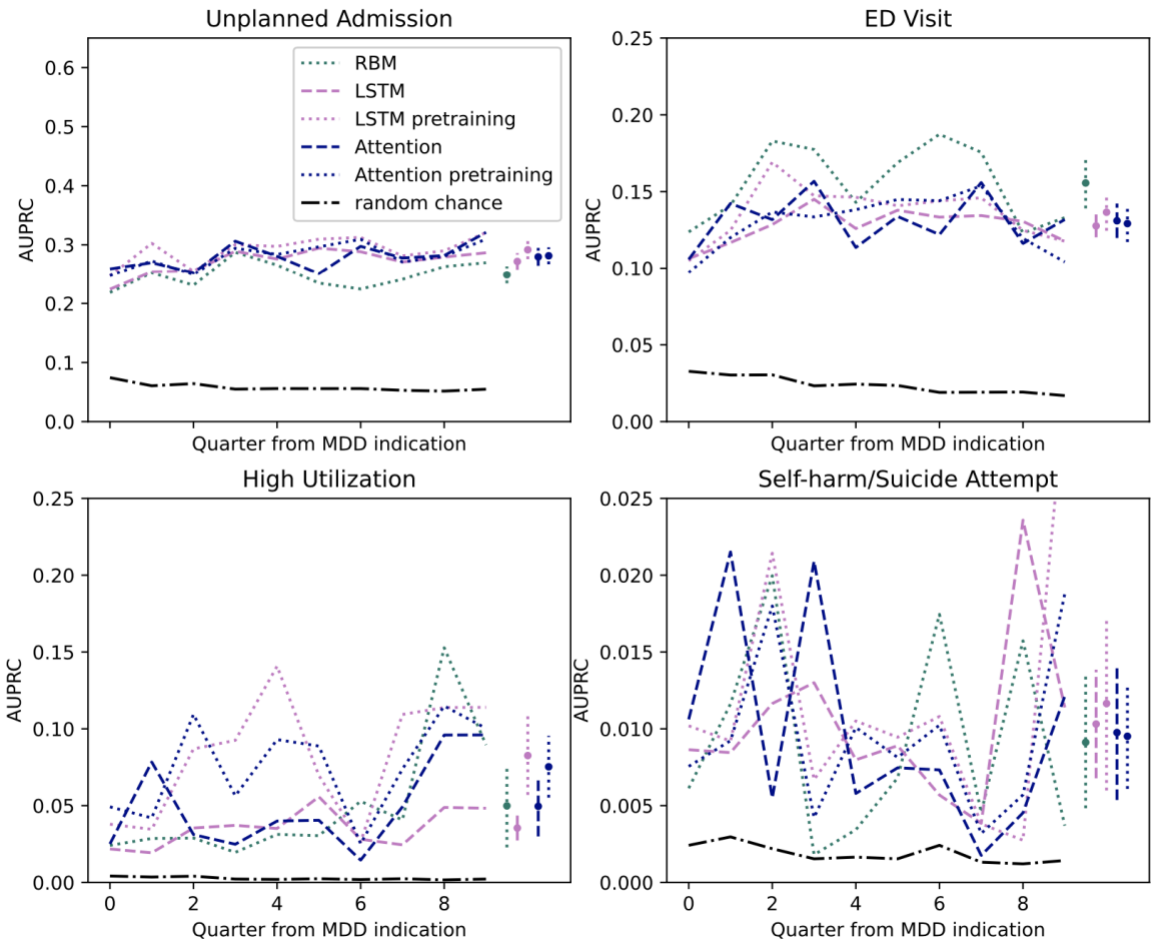


Figure 1-4 Temporal validation of pretraining methods.

LSTM and Attention pretrained models were compared to the same model structure without pretraining. The RBM model was best performing of feature engineering methods and is included in this graphic. The black dot-dash line represents the performance of a random chance estimator. 95% confidence intervals are shown for the average AUPRC across the study time period.

Discussion

This study evaluated autoencoder feature engineering and pretraining in MDD patients across an array of prediction tasks. Autoencoder models selected for this study account for temporality, denoise, reduce dimensionality and capture interactions between EHR features. When using pretrained weights we were able to improve predictive performance over benchmarks in three of four outcomes. The pretrained weights improved predictive performance in the LSTM, relative to a model with the same architecture and no pretraining. This suggests that pretrained information from the autoencoders may be best retained in the model weights. In contrast, encodings as input to a random forest model did not improve predictive performance. Improvement in prediction resulting from pretraining has potential for increased clinical usefulness of risk models in MDD and other clinical areas, with a test precision in the self-harm/suicide attempt outcome over 1%.[8,29,51]

Feature encodings from LSTM and Attention autoencoders were not superior in predictive performance relative to aggregate feature engineering. Random forest models have a decision tree structure that accounts for feature interactions and ensembles decision trees to protect from overfitting to noise in the training data.[52] It appears that the random forest model's structure was sufficient to partially account for noise and complex interactions while the benefit of temporality captured in encodings in the LSTM and Attention varied across prediction tasks. We observed information loss in training each autoencoder format, as none were able to achieve zero validation loss. The information lost in encoding estimation may contribute to the lack of performance of autoencoder feature engineering, while this information may have been retained in autoencoder weights.

Encodings from the RBM model resulted in best AUPRC for all but one outcome of the feature engineering techniques. The RBM modeling strategy used has been shown to outperform principal component analysis as a feature engineering technique in a suicide risk prediction task and has the capability to learn complex interactions

between high dimensional features. Additionally, the RBM has a low number of trainable parameters (29,747 to 74,147) relative to the LSTM and Attention autoencoders, suggesting LSTM and Attention autoencoders could be overfit.

We observed variation in performance between autoencoder configurations with the LSTM pretraining model having the most consistent performance across prediction tasks. Attention based models have outperformed LSTMs in many sequential data prediction tasks, specifically in natural language processing.[26,48,53,54] Attention models are effective in part because of their ability to efficiently learn long range dependencies in a sequence. However, in our study the sequences are relatively short compared to NLP applications where attention-based models have been superior. Additionally, the structure of attention models allows for increased parallelization—speeding up training relative to LSTMs. In our study the model trainable parameters and number of training examples were such that training time was relatively short. The nature of this study may nullify the advantages attention models have over LSTMs in other studies.

Average AUPRCs were low for the self-harm/suicide attempt outcome across prediction techniques. Self-harm/suicide attempt events were relatively rare compared to the other outcomes, except for high utilization. The high utilization outcome has similar frequency, but pretrained LSTM has a mean AUPRC more than seven times that of self-harm/suicide attempt. Since EHR data reflect healthcare utilization, EHR based features may provide higher prediction performance in utilization-based outcomes. Overall, healthcare utilization is common in patients prior to suicide events, although not always mental health care-specific utilization.[55] Further study of health utilization events that precede suicidality on the causal pathway could allow for training of prediction models with increased clinical precision.

This study highlights several implications for the use of autoencoders for prediction tasks in the MDD population. Evidence of the benefits for autoencoder pretraining is shown with a limited dataset at a single site. Researchers considering development of predictive models in this patient population may improve predictive performance with this training strategy. It is possible the benefits of autoencoder pretraining will extend to additional clinical areas.[24] LSTM performance relative to Attention architectures suggests that LSTM architectures should also be considered

when working with similar datasets. Observed autoencoder information loss, specifically in Attention architectures, could have been due to the lack of training examples. Future studies of multi-site data may have better performance in Attention architectures and allow for additional training techniques such as self-supervised learning.[26] We observed higher AUPRCs in health utilization outcomes relative to self-harm and suicide attempts. Researchers in this space should consider the actionability and clinical usefulness of these or related health utilization outcomes when developing risk models for MDD patients.

This study is limited to a single site and single mental health phenotype. It is possible that the study results will not generalize to other phenotypes and health care systems. Use of billing codes for outcomes, specifically self-harm/suicide attempt, may not capture all cases, potentially biasing results.[56] We study multiple autoencoder structures of varying model sizes. However, there are many alternative structures not studied here that may result in improved performance.

Conclusion

We evaluate temporal autoencoder pretraining and feature engineering in the MDD population and compare predictive performance to a benchmark modeling strategies that have proven successful in the MDD phenotype.[12,33] LSTM models with pretrained weights from autoencoders were able to outperform the benchmark, as well as an equivalent LSTM model without pretraining. Autoencoder feature engineering was unable to outperform the benchmark. This suggests that information retained by model weights may not be passed to encodings. Future researchers developing risk models in MDD may benefit from the use of autoencoder pretrained weights.

Chapter 2 Empirical Calibration of Antidepressant Adverse Events in Observational Electronic Health Record Data

Background and Significance

Major Depressive Disorder (MDD) is a prevalent and heterogeneous phenotype with a variety of treatment options available. Many challenges hinder optimal treatment decisions for individuals with this disease as comparative efficacy and safety may be unclear.[4,57] The challenge may be exacerbated by the fact that most individuals providing care are not mental health specialists.[1] Switching antidepressant medications or using adjunctive antidepressant therapy is commonly done in the course of treatment and may be due to either adverse events or non-response to the original medication.[57] The safety of such practice, compared to continuing the initial treatment or terminating treatment altogether, is an important consideration in treatment decisions.[58,59] The STAR*D study, a large pragmatic clinical trial of MDD treatment, found an increased risk of adverse events in patients who switched to a different antidepressant from an initial prescription of Citalopram.[13] In the OPTIMUM study the safety and efficacy of augmentation against switching strategies were evaluated in geriatric patients with treatment resistant depression. Variation in adverse events were observed between treatment strategies with an increased fall rate in patients treated with bupropion augmentation.[14]

Electronic health record (EHR) data allow for the investigation of efficacy and safety of clinical practice; however, harms are more commonly recorded in medical records, rendering studies of adverse events more feasible.[60] EHR data are a useful source for adverse event research, as they can capture a large and diverse population of patients who receive antidepressant therapy in routine care, reflect current prescribing patterns and practices, and allow for long-term follow-up and outcome assessment.[61] Because of these attributes of EHR data, the United States Food and Drug Administration (FDA) has provided guidance on the use EHR and claims data in the generation of real world evidence (RWE) to improve regulatory decisions.[62] In their guidance, the FDA encourages the use of RWE for clinical trial hypothesis generation as well as post market safety and efficacy surveillance.

A challenge of this work is the observational nature and limited data available in EHRs. If certain untestable causal inference assumptions[63] are not met, estimates of intervention effects will be biased. Many standard causal estimation procedures assume conditional exchangeability.[64] Conditional exchangeability—the assumption that common causes of treatment and outcomes are measured—is of primary concern in an observational study, and can

lead to systematic bias in effect estimates. There is no empirical method to prove that the exchangeability assumption is met, leading many to a justifiable suspicion of the results of observational studies. Recent studies[15,60,65–68] have applied control outcomes as a means of correcting for the systematic biases that often arise in observational studies. Empirical calibration with control outcomes has the capability to correct for confounding, selection bias and measurement error.[66] In order to correct for these biases, control outcomes should come from a diverse sample that share the mechanisms of bias with the study outcomes. To accomplish this requires an understanding of confounding and other mechanisms of bias between the treatment of interest and the control and study outcomes.[69] In a high-throughput replication study of the Observational Health Data Science and Informatics (OHDSI) research network, it was found that observational studies may overestimate significance due in part to systematic bias.[60] Negative control outcomes were used to show that the impact of residual confounding is minimal in an observational study of Oseltamivir—an antiviral influenza treatment—and complications.[70] Empirical calibration with control outcomes has been used to calibrate treatment effects in studies of vaccine efficacy[71] as well as the efficacy and safety of hypertension treatments.[67]

Objectives

In this study, we evaluated adverse event risk of prescribing a new selective serotonin reuptake inhibitor (SSRI) or serotonin and norepinephrine reuptake inhibitor (SNRI) antidepressant in patients with a history of antidepressant therapy. The adverse events include insomnia and nausea, which are commonly reported side effects in antidepressants.[1,4,57] For each outcome, we evaluate covariate matching and inverse probability of treatment weighting (IPTW), two commonly used causal inference methods, to estimate treatment effects in a patient population with history of antidepressant therapy at VUMC. Covariate matching alters the population about which inference is made to the treated population, often referred to as the average treatment effect in the treated (ATT). A potential benefit of matching analysis in this observational setting is that the population is in part informed by clinical practice, as the matched controls, by construction, have similar covariate distribution to their treated counterparts. IPTW is used to estimate the average treatment effect (ATE) for the entire population.

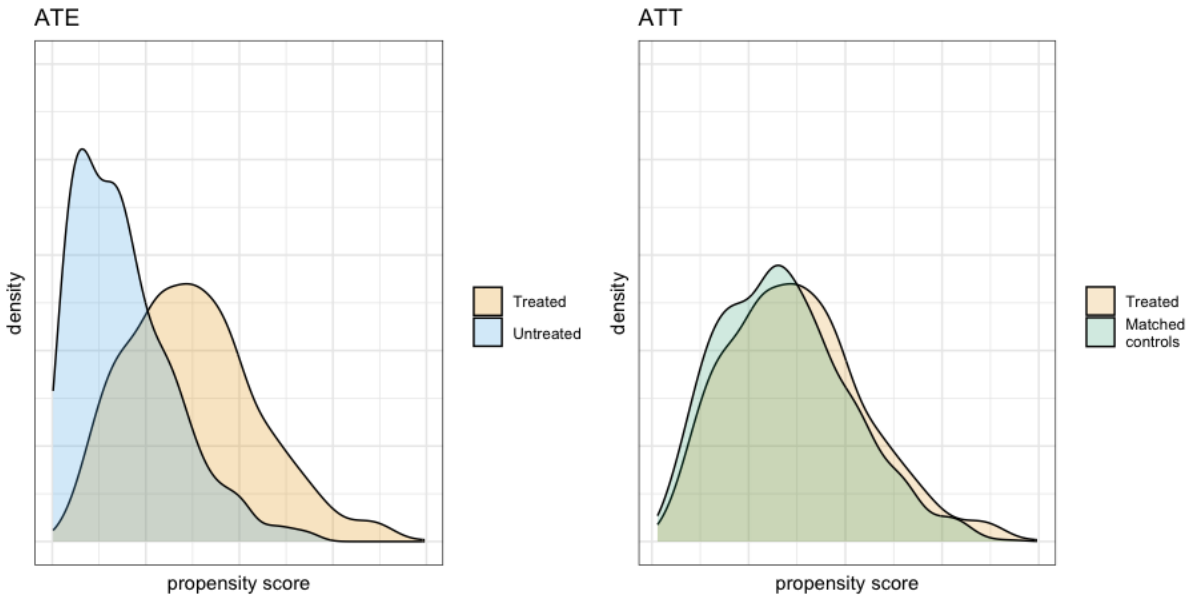


Figure 2-1 Illustration of the difference between the populations about which inference is made when calculating ATE vs. ATT.

The ATE is an estimate of the treatment effect in the entire study population. In many observational studies the treated and untreated populations may have differing covariate distributions, which can be reflected by their propensity scores as shown above. IPTW attempts to balance the distributions of observed covariates between the treatment and control groups by weighting. In contrast, ATT estimates the treatment effect in the treated population by matching treated to untreated patients with similar covariates. This results in a balancing of covariates between the treated and matched population and equipoise in the propensity score distributions.

We investigated bias and calibrated estimates with empirical calibration of control outcomes. We selected a set of negative control outcomes through a semi-automated process. Negative controls and synthetically generated positive controls were used to estimate an error model that quantifies bias and calibrates confidence intervals.

Methods

Data Description

This study was conducted using data from the Vanderbilt University Medical Center (VUMC) Research Derivative,[41] an identified secondary-use research data warehouse. VUMC is an academic medical center based in Nashville, Tennessee, in the United States Mid-South. We sought to identify patients with a history of MDD

treatment recorded at VUMC, specifically having likely undergone a prior antidepressant trial. Patients were included in the study if they had a visit to VUMC between 1/1/2005-12/31/2021, were prescribed an antidepressant and a depression-related International Classification of Diseases (ICD) diagnosis code, including MDD, dysthymic disorder, depressive disorder not elsewhere classified,[42,43] or problem list mention of depression within 56 days following the prescription. The ICD codes are included in appendix table A.1 and the antidepressant list in table A.2. Patients were also required to undergo 56 days without being prescribed a different antidepressant than initially prescribed, as the recommended time for an antidepressant trial is 6-8 weeks.[57] We excluded patients diagnosed with bipolar and schizophrenia disorder by ICD diagnosis codes included in appendix Table A.12.

For patients meeting entry criteria, each patient's record is divided into evaluation periods. The evaluation period is treated as the unit of analysis in this study, where an individual patient may have multiple evaluation periods included in the study. The first evaluation period for a given patient starts at their first visit after meeting entry criteria. The evaluation period is further divided into an intervention and outcome period. The intervention period starts at the beginning of the visit and continues until the visit end or an intervention that takes place within 14 days of the end of the visit, whichever is later. The outcome period is the 180 days following the end of the intervention period. The next evaluation period begins at the first visit after the 180-day outcome period.

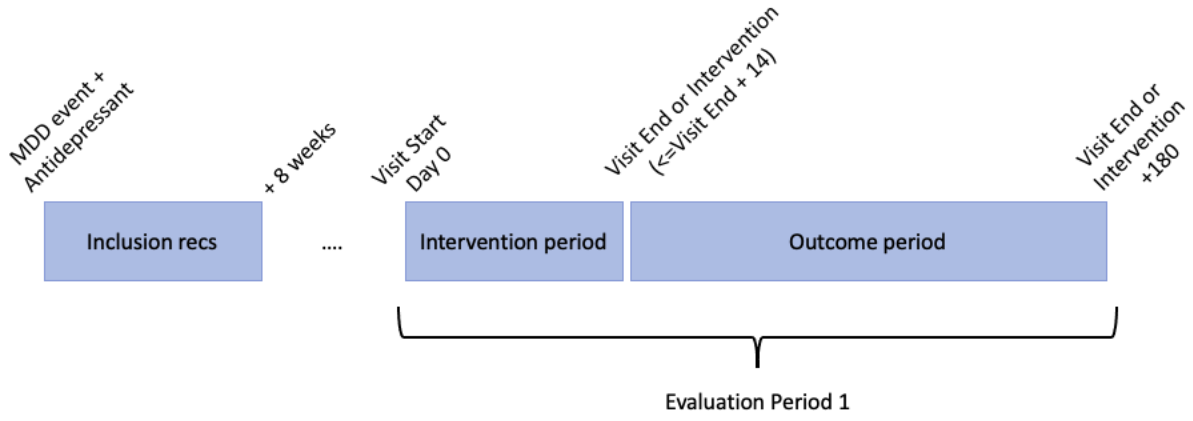


Figure 2-2 Visualization of the evaluation period definition.

Evaluation periods are divided into intervention and outcome periods. The first episode begins at the start of the first patient visit once inclusion criteria is met. The intervention period goes from the start of the visit until the visit

end, unless the patient is prescribed an antidepressant. In that case the intervention period is ended on the day the intervention takes place. The outcome period is the 180 days following the intervention period.

The interventions compared in this study are prescribing a new SSRI or SNRI antidepressant and no antidepressant prescribed during the intervention period. New prescriptions are those for which the patient has not been prescribed the drug in the prior 180 days, where drugs are identified at the Anatomical Therapeutic Chemical (ATC) level 5.

To reduce risk of spurious results from multiple analyses, the adverse events considered in this study were limited to those deemed high priority in antidepressant prescribing by clinical experts (CGW) and were observed frequently enough to power our study (see appendix table A.8). There are many other adverse events that are considered in antidepressant prescribing,[72] however we focus our analysis on insomnia and nausea, defined by SNOMED CT codes from a recent study.[60] Outcome inclusion and exclusion codes are provided in appendix Table A.5. For the effect estimates of each outcome, patients with a prior instance of that outcome in their medical record were excluded.

The confounding mechanisms between the treatment and study outcomes are widely studied, but not fully understood. We hypothesize that factors such as MDD severity, treatment resistance, socioeconomic status, social support, and comorbid conditions are likely influential.[34,73–75] These variables may not be well represented in the EHR. However, we include covariates that may represent proxies of these variables and may substantially reduce (although not fully eliminate) the contaminating impact of confounding.[64] Empirical calibration provides insight into the degree to which confounding bias is controlled. Study covariates include age, gender, and race/ethnicity, insurance status, health system utilization, MDD interventions, antidepressant prescribing history and comorbidities. Comorbidities were formatted using Agency for Healthcare Research and Quality Clinical Classification Software (CCS)—a grouping of ICD codes intended to be clinically meaningful.[44]

Risk modeling

Average treatment effects on the treated (ATTs) were estimated by k:1 covariate matching. Inclusion of additional controls through k:1 matching has potential to increase study power, but may sacrifice covariate balance. The selection of k was done in consideration of covariate balance and power. Details of this analysis are included in Table A.8 of the appendix. We compared patients prescribed an SSRI or SNRI to those with no prescription.

Logistic propensity score matching was used to create a subpopulation composed of treated cases and matched controls from the untreated. This dataset was used to make inference about the treated population. Inverse probability of treatment weighting (IPTW) was used to estimate average treatment effects (ATE). Propensity scores were estimated by logistic regression and IPTW weights were truncated at a maximum of 5.[76,77] Standard errors were estimated using a sandwich cluster-robust covariance matrix, clustered by patient.[78] Matching and weighting were done separately for each outcome excluding evaluation periods with a prior instance of the outcome.

Selection and synthesis of control outcomes

Candidate negative control outcomes were selected by accounting for drug product labels, FDA adverse event reporting, publications and observed frequency in study population as in Schuemie et al.[66] For inclusion in the study, we required negative control outcomes to have case frequency in the treated population in a range similar to the study outcomes. To generate the recommended 30 or more[66] negative control outcomes, we grouped concepts based on common ancestor, iteratively selecting the grouping with highest number of cases and least number of concepts. Once a concept was grouped, it was excluded from future groupings. We continued this process until all concepts were part of a grouping. Groupings were reviewed and those considered too broad were excluded. We then reviewed all negative control outcomes for those likely to share confounders with the study outcomes. For those negative control outcomes with likely shared confounding, we evaluate calibration effectiveness using the remainder of the negative control outcomes. This provides insight into calibration effectiveness on an outcome for which the true risk ratio is likely known and potentially shares confounders with the study outcomes.

We used negative controls to synthesize positive controls for confidence interval (CI) calibration. The relationship between treatments and negative controls may exhibit confounding, if this confounding not accounted for in positive control generation results could be overly optimistic.[66] We attempted to preserve measured confounding by fitting a logistic regression model with the negative control as dependent variable and covariates and independent variables. This regression model learned associations between the covariates and negative control. The estimated probability from the logistic regression was used to simulate new cases in the treated population from a Bernoulli distribution. If there were n cases in the treated population, we simulated an additional n cases resulting in a risk ratio of 2.0. For a risk ratio of 1.5, we simulated $0.5n$ cases and for a risk ratio of 3.0, we simulated $2n$ cases. Negative controls can

represent both unmeasured and measured confounding, however positive controls only accounted for measured confounding represented by the logistic regression model.[66]

Empirical Calibration

Empirical calibration was done using the EmpiricalCalibration R package.[15,66] The control outcomes were used to fit an error model that accounted for systematic bias in our study. This approach assumes that the error follows a normal distribution about the true effect and that the mean and variance of the error distribution were assumed to be linear functions of true effect size, where the slope and intercept of these functions is estimated by likelihood maximization. The resulting systematic bias model provided a description of observed bias, in both mean and variance, the intercept of the linear model measuring bias for a null effect, and the slope estimated the relationship between bias and true effect magnitude. Additionally, this model assumes that treatment effects on the study safety outcomes follow the same bias distribution as the negative controls. The resulting error model to calibrates point estimates and standard errors to account for systematic bias when assumptions are met.

Results

The full study population includes 69,298 patients with 233,680 evaluation periods, 81% of patients had multiple evaluation periods and the median number of evaluation periods per patient was 3. Analysis of power and covariate balance of matching strategies resulted in 4:1 matching. See the appendix Table A.8 and A.9 for further details on the matching analysis. Table 1 displays the full and matched populations without prior outcome exclusion. The matched population has 17,120 patients and 19,525 evaluation periods, which accounts for 8.4% of evaluation periods in the full population. Table 1 includes demographic variables and covariates that deviate in distribution between the full and matched populations. The matched population is younger on average and has elevated rates of unplanned admissions and ED visits. The matched population has higher rates of prior comorbidities including suicidal ideation/behavior, mood disorders, and anxiety disorders. The full and matched populations have the same number of treated evaluation periods, 4,007.

Table 2-1 Study summary statistics for full and matched population. Included are demographics, comorbidity rates for a subset that deviate in frequency between full and matched populations, and outcome rates.

		Full population	Matched population
	N Evaluation periods (Unique Patients)	233,680 (69,298)	19,525 (17,120)
	Median Age (IQR)	52.9 (25.7)	48.1 (17.1)
		N (%)	
	Female	165,356 (70.1%)	13,584 (69.6%)
Race/Ethnicity	**Asian**	2,234 (1.0%)	151 (0.8%)
	Black	21,515 (9.4%)	1,683 (8.6%)
	White	197,115 (85.9%)	16,877 (86.5%)
	Hispanic-Latinx	2,837 (1.2%)	331 (1.7%)
	Other	5,877 (2.6%)	483 (2.5%)
Prior healthcare utilization	**Unplanned admission**	9,829 (4.3%)	5,865 (30.0%)
	ED visit	4,856 (2.1%)	2,596 (13.3%)
Prior MDD Interventions	**Psychotherapy referral**	4,375 (1.9%)	1,805 (9.2%)
Select Prior CCS Comorbidities	**Suicidal Ideation/Behavior**	4,175 (1.8%)	991 (5.1%)
	Mood disorders	120,969 (52.7%)	13,041 (66.8%)
	Anxiety disorders	65,304 (28.4%)	8,329 (42.7%)
	Disorders of lipid metabolism	70,916 (30.8%)	4,603 (23.6%)
	Eye disorders	36,266 (15.8%)	1,973 (10.1%)
Payor	**Medicaid**	6,665 (2.9%)	1,020 (5.2%)
	Selfpay	5,360 (2.3%)	774 (4.0%)
Safety Outcomes	**Insomnia**	3,342 (1.5%)	326 (1.7%)
	Nausea	7,169 (3.1%)	614 (3.0%)

A total of 42 negative controls met selection criteria. Of those 15 did not require grouping (see appendix Table A.6) and 27 were grouped (see appendix Table A.7). The psycho-active substance abuse negative control is likely to share some confounding structure with the study outcomes as MDD has been shown to be a risk factor for psychoactive substance abuse and is commonly associated with MDD.[79–82] Because psychoactive substance abuse is likely to share confounding with the study outcomes and we do not identify evidence for a causal relationship between the study treatments and this control outcome, we evaluated calibration of psychoactive substance abuse using the remaining 41 control outcomes in the full population. This provides insight into the effectiveness of calibration in an outcome with known risk and potential shared confounding with the effects of interest in our study.

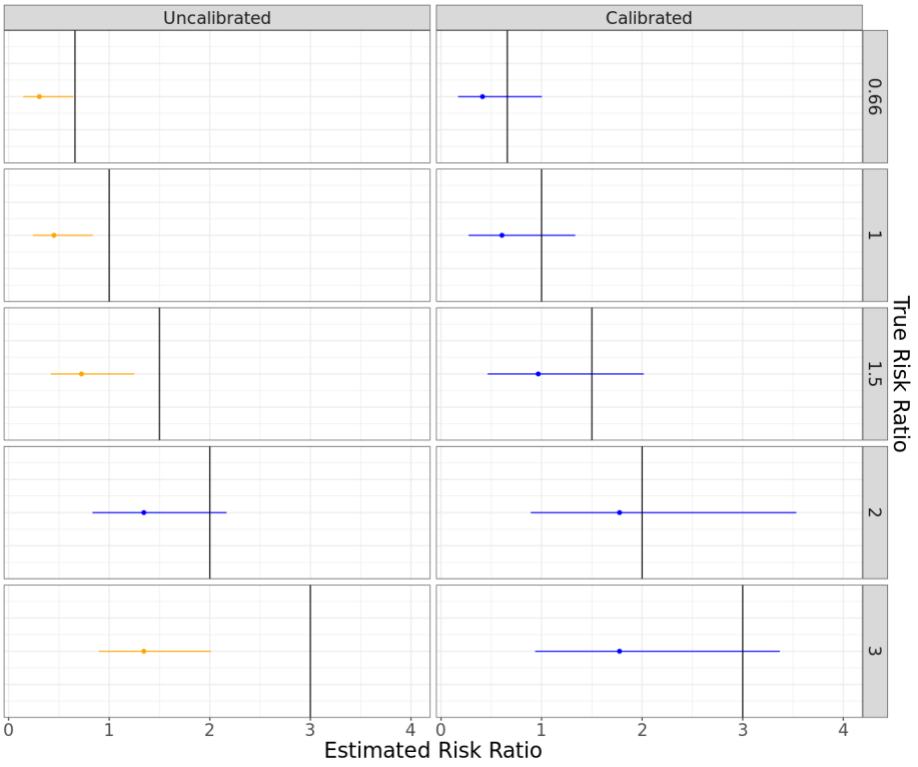


Figure 2-3 Calibration of psychoactive substance abuse control outcome forest plots in the full population. The left panel shows uncalibrated estimates, the right estimates after calibration by remaining control outcomes. Confidence intervals that capture the true value are blue and those that do not are orange.

Uncalibrated estimates for psychoactive substance abuse do not include the true risk ratio at four out of five levels including risk ratio equal to 1—where no alteration of the control outcome took place. Calibration results in all five

confidence intervals capturing the true risk ratio, while point estimates are increased by calibration, the bias may not be fully corrected, as the point estimate is less than the true risk ratio at each level.

The negative control calibration model for ATEs had an intercept of -0.29 with slope of 1.0. Suggestive of a protective bias consistent at each risk ratio. The calibration of standard errors had an intercept of 0.26, suggesting an underestimation of variance.

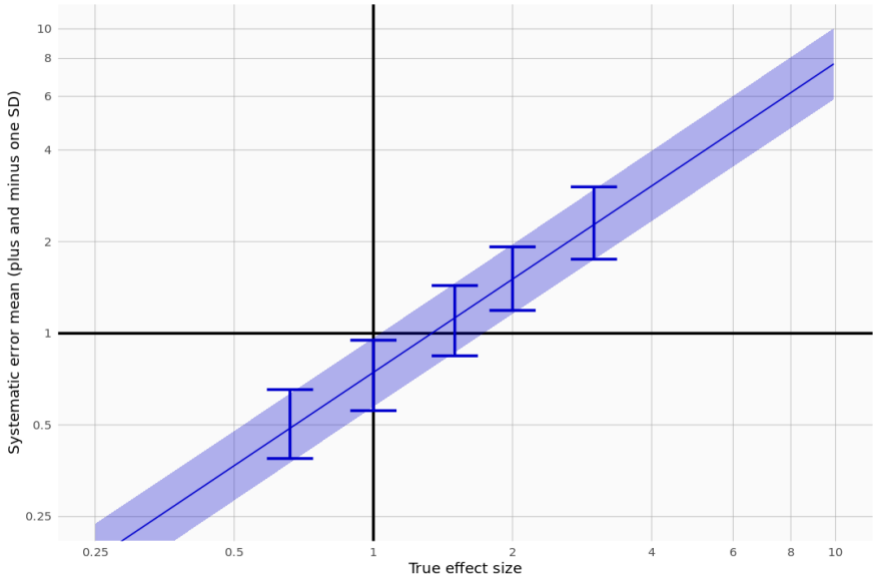


Figure 2-4 Plot of ATE error model.

The x-axis represents the true risk ratio and y-axis the model estimation of systematic error. Estimates of mean error are shown by the dark blue line with plus or minus one standard deviation shading. Error bars indicated plus and minus one standard deviation at each risk ratio level.

The calibration model for ATTs had an intercept 0.05 and slope 0.99. The intercept provides evidence that risk may be overestimated in ATT. The calibration of standard deviations had an intercept of 0.07, suggesting an underestimation of standard errors.

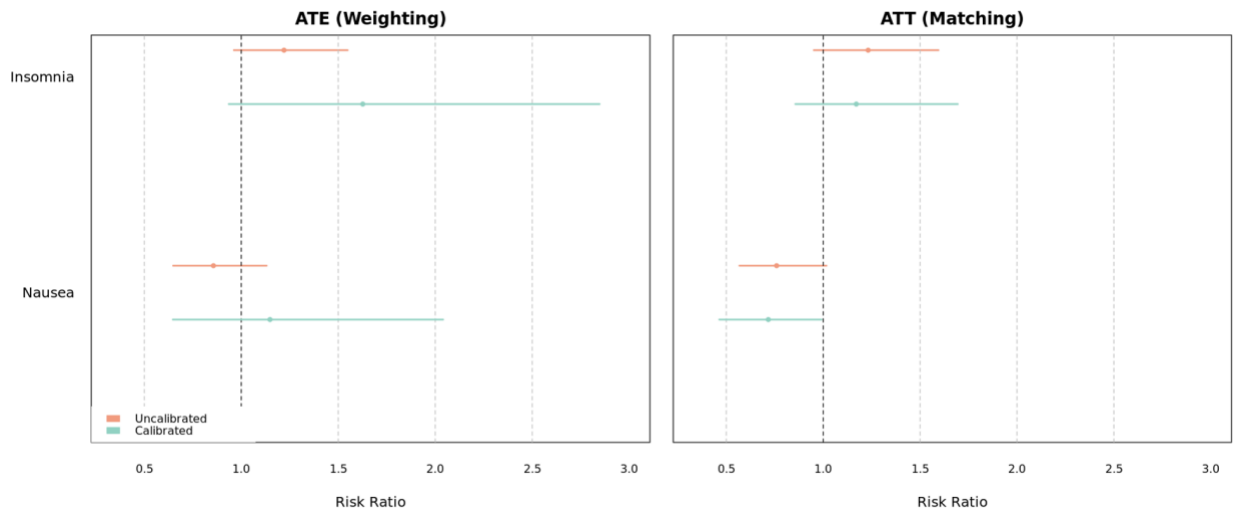


Figure 2-5 Forest plot of ATE and ATT estimates.

The ATE (left) and ATT (right) estimates with (green) and without (orange) empirical calibration

Due to potential protective bias and underestimation of standard errors in ATE, point estimates and standard errors were increased by calibration. The uncalibrated ATE risk ratio estimate for Nausea is 0.87 (95% CI 0.67-1.15), with calibration the estimate is 1.18 (95% CI = 0.66-2.08). ATT estimates are adjusted in the protective direction due to a potential increased risk bias. The ATT estimate for Nausea is 0.76 (95% CI = 0.57-1.02), and with calibration 0.72 (95% CI = 0.46-0.99).

Discussion

In our observational investigation of adverse event risk of antidepressant switching, we observed a strong protective bias in ATE estimates and underestimation of standard errors. In ATT estimates we observed a reduced and inverse—increased risk of—bias. The ATE and ATT estimate risk in different patient populations, potentially leading to different biases. It is possible that a selection bias exists in the full population, that is reduced in the matched. The calibrated ATT estimates are also more precise than the ATE, even though reducing sample size to estimate ATT decreases power. This may be due to variation in biases between negative controls in the ATE estimation scenario, which necessitates an error model that requires large corrections to standard errors in order to account for these biases.

The ATT estimates are more robust in this study and may be higher yield for generation of clinical evidence in observational studies of this phenotype. ATT estimates may benefit from clinical expertise informing selection of the study population, as the controls are selected to resemble the covariate distribution of those selected by clinicians for treatment. We find that after calibration there is evidence for a protective effect in the Nausea outcome. This finding may be unexpected due to the common observation of antidepressant discontinuation due to Nausea.[83,84] However, we limit our study population to those with a history of antidepressant treatment, it may be that for those who have a record of prior tolerability to antidepressant therapy that the study intervention is protective. Prior studies have shown a decreased rate of adverse event discontinuation in patients switching antidepressants relative to antidepressant initiation.[58] We also should take into consideration the potential for under reporting or documentation of Nausea cases, as we observe Nausea cases at a lower rate than studies of antidepressant side effects with active follow up.[72]

When protective biases are observed in observational studies, as in the ATE estimates, it is often a case of healthy user bias.[85] It is possible we observe a related form of bias due to selective prescribing. Gill et al.[16] show that patients with multiple comorbidities are less likely to be prescribed an antidepressant. In our study the treated population has fewer CCS codes on average (9.0 treated vs 9.7 untreated). It is possible we are observing a selective prescribing to patients with lower comorbid burden, and the lower comorbid burden population may be less likely to develop the negative control outcomes in the future. It is also possible other forms of bias are impacting this result such as unobserved confounders.

This study is limited by insufficient power to investigate SISB risk. To achieve sufficient power in the matching analysis we performed k:1 matching, which sacrificed covariate balance—introducing bias into the ATT. We also lack power to investigate switching from and to different medications. The grouping of negative control outcomes may have introduced additional noise into calibration. Despite these limitations, our study provides valuable insights into the safety of prescribing practices and highlights the potential for systematic monitoring of adverse events. Furthermore, the ATT population shows limited evidence of bias and has potential for further analysis including heterogeneity. The study methods have the potential to be applied in other settings to monitor prescribing practices and identify potential safety concerns.

Conclusion

Our observational investigation of adverse event risk in patients prescribed a new antidepressant revealed evidence for a protective bias in ATE estimates. The protective bias observed may be due to selective prescribing, or another form of bias. We observed a reduction in bias in ATT estimates. It is possible that a selection bias exists in the ATE population, which is reduced or eliminated in ATT estimates. The methods studied can be used to monitor and improve safe prescribing practices. Furthermore, the ATT population shows limited evidence of bias and is suitable for further analyses, including heterogeneity, which we will address in Chapter 3. While our study is limited by its observational nature, it provides valuable insights into understanding and correcting biases when estimating antidepressant adverse event risk.

Chapter 3 Machine learning estimation of antidepressant adverse event heterogeneity

Background and Significance

Major Depressive Disorder (MDD) is a complex disease that is challenging to treat and manage.[4,57] Treatment decisions are complicated by the fact that the majority of antidepressant prescriptions (approximately 79%) are written by providers who do not specialize in mental health care.[1] It is common practice to switch antidepressants due to non-response or adverse events.[57,86] The safety of such practice, compared to continuing the initial treatment or terminating treatment altogether, is an important consideration in treatment decisions.[13,14,58,59]

Further, heterogeneity of antidepressant adverse events in MDD subpopulations has been evaluated in recent studies. The OPTIMUM study focused specifically on the geriatric-treatment resistant depression subpopulation.[14] Gill et al. measured variation in antidepressant prescribing practices given comorbid burden and found that patients with multiple comorbidities are less likely to be prescribed antidepressants.[16] Elderly patients are more likely to be multimorbid and have declined drug metabolism, leading to an increased risk of adverse effects.[4] Köhler-Forsberg et al.[17] identified adverse event and discontinuation heterogeneity by comorbidity, noting an increased risk in patients with comorbid fibromyalgia and neuropathic pain. Inconclusive results have been found in studies of MDD and comorbid cancer.[18] Efficacy in studies of patients with comorbid MDD and alcohol dependence found mixed results dependent on selected outcomes.[19]

Statistical analysis of heterogeneity is often under-powered and can lead to spurious findings when multiple subgroups are analyzed; because of this, clinical trials often require pre-specification of subgroups,[87] limiting investigators' ability to discover unforeseen sources of heterogeneity.[88] Heterogeneous Treatment Effect (HTE) models reframe the problem to focus on the detection of heterogeneity, rather than conducting a stepwise subgroup analysis. This shift has been facilitated by machine learning methods, which allow for flexible modeling of high dimensional data. HTE models are machine learning models designed to estimate the conditional average treatment effect (CATE). The average treatment effect (ATE) refers to the difference in mean potential outcome that compares hypothetical settings in which each treatment is applied to the entire population.[63] The CATE, on the other hand,

estimates the ATE conditional on covariates, and can therefore provide insights into treatment effect variation by patient characteristics.[89] Common machine learning models such as random forests and gradient boosting have been adapted for CATE estimation.[88,90]

Electronic Health Records (EHRs) are a valuable resource for research on real-world treatment practices, particularly in the field of adverse event research.[60] These records provide a comprehensive view of a diverse patient population who are undergoing antidepressant therapy in routine care, thereby reflecting current prescribing patterns and practices.[16] Moreover, EHRs enable long-term follow-up and outcome assessment. The Food and Drug Administration (FDA) has placed emphasis on the importance of real-world evidence in making regulatory decisions.[62] However, the observational nature of EHRs and the limited data they contain pose difficulties. If certain untestable causal inference assumptions are not met, the resulting estimates of intervention effects may be biased.[91] Recent studies offer a technique to measure systematic biases often encountered in observational studies using negative controls and semi-synthetic positive control outcomes.[65–67]

Semi-synthetic outcomes are generated using the features and outcomes as a baseline and then synthetically altering the outcome according to a specified data generating function. The utilization of semi-synthetic outcomes presents a valuable approach in analysis of heterogeneity for several reasons. First, true CATE at the individual level is unknown, even in labeled data,[92] because calculation of a treatment effect at an individual level requires knowledge of the counterfactual scenario where the patient receives each treatment considered.[93] Secondly, real-world data may contain unaccounted for and unmeasured sources of bias.[15] Also, semi-synthetic outcomes allow for evaluation of model performance under varying conditions, such as changes in noise and feature density. This can offer researchers insights into the underlying characteristics in real data. While numerous studies have been conducted on synthetic or semi-synthetic data using benchmarking datasets,[90,92,94] in this study, we evaluate HTE model performance on a real-world dataset with semi-synthetic outcomes followed by analysis of non-synthetic adverse event outcomes. The semi-synthetic outcomes, selected using the control outcome strategy of Schuemie et al.,[66] allow us to analyze data with a known CATE generated from a real feature set across outcomes with varying relationships to features. The comparison of semi-synthetic to real outcome performance prevents drawing conclusions based solely on synthetically generated data, which may favor one algorithm over another due to data

generating process alone.[92] Instead, the results of the semi-synthetic analysis inform our evaluation of model performance on non-synthetic outcomes, potentially enhancing the interpretability of our findings.

Objectives

This study aimed to evaluate HTE models and their performance under varying data generating processes with semi-synthetic outcomes in an MDD patient population. We also examined evidence for heterogeneous treatment effects of SSRI/SNRI prescribing on the occurrence of adverse events—insomnia and nausea. Another key objective is to identify which model is best calibrated to estimate this heterogeneity and what this implies about the underlying data. Through this research, we sought to enhance understanding of HTE models and their potential to personalize the antidepressant prescribing processes.

Methods

Data Description

The study data were extracted from the Vanderbilt University Medical Center (VUMC) Research Derivative, an identified research data warehouse.[41] We studied patients likely to have undergone an antidepressant trial and had a visit between 1/1/2005 and 12/31/2021. Inclusion required patients were prescribed an antidepressant (see Table A.2) followed by a 56-day period in which no antidepressants were prescribed, as the recommended time for an antidepressant trial is 42-56 days.[57] We also required patients to have either a problem list mention of MDD, or an MDD related ICD diagnosis code within 56 days following the initial antidepressant prescription. The list of ICD diagnosis codes is included in appendix Table A.1 and include MDD, dysthymic disorder, depressive disorder not elsewhere classified. We excluded any patients diagnosed with bipolar and schizophrenia disorder by ICD diagnosis code; exclusion codes are provided in Table A.12.

For patients that met entry criteria, medical records were segmented into what we refer to as evaluation periods. These evaluation periods serve as the unit of analysis, and a patient may have several evaluation periods included in the study. A patient's first evaluation period commences with their initial appointment after they meet the entry

requirements. This evaluation period is then further split into two distinct phases: the intervention period and the outcome period. The intervention period begins at the start of the visit and lasts until either the end of the visit or at the time of any intervention that occurs within 14 days of the visit's conclusion, whichever took place later. The outcome period includes the 180 days that follow the end of the intervention period. The subsequent evaluation period is initiated at the patient's first visit following the 180-day outcome period.

This study compares new SSRI or SNRI antidepressant prescriptions with no new prescriptions or unchanged pharmacotherapy. New prescriptions are prescribed drugs with no record of that drug in the prior 180 days, identified at Anatomical Therapeutic Chemical (ATC) level 5. The adverse events considered in this study are the same as in chapter 2, SNOMED CT code definition are included in appendix Table A.5. Patients with previous instances of the outcome were excluded from effect estimates.

The relationship between treatment and study outcomes is complex and not fully understood. Factors such as MDD severity, treatment resistance, socioeconomic status, social support, and comorbid conditions are hypothesized to be influential.[34,73–75] Our study includes proxy variables for these factors to reduce confounding impact.[64] Study covariates include age, gender, race/ethnicity, insurance status, healthcare usage, MDD interventions, antidepressant history and comorbidities. Comorbidities were grouped using Agency for Healthcare Research and Quality Clinical Classification Software (CCS)—an ICD code grouping intended to be clinically meaningful.[44]

We showed in chapter 2 that covariate matching may reduce bias in treatment effect estimates. In order to potentially reduce bias and increase ability to identify heterogeneity, we restrict our analysis to a covariate-matched population. Given that the control population is selected to have similar baseline characteristics to the treated population, our target estimand therefore is more appropriate interpreted as a conditional average treatment effect among the treated (CATT).

HTE model descriptions

We evaluate HTE models leveraging the double machine learning (DML) paradigm from Athey and Imbens.[89] DML models assume the data generating process $Y = \theta(X) \times T + g(X) + \epsilon$, with $T = f(X) + \eta$ where g and f can be

any machine learning model, θ is the final function of heterogeneity and ϵ and η are error terms. The causal forest model applies a modified random forest in which the tree splitting criteria maximizes the variance of $\tau(X)$ between sub-trees.[88] Causal forest parameter estimates are shown to be asymptotically normal, allowing for valid and efficient CI estimation under sufficiently large samples. Additionally, causal forests have been shown to perform well in scenarios where there is low heterogeneity in treatment effect.[92] The gradient boosted DML has shown superior performance to causal forest in some simulation studies, but is a higher variance estimator that may be prone to overfitting.[90,94] We also include Linear DML models in this study that assume linearity in θ, and have a reduced risk of overfitting. As a benchmark model, we fit a random forest to each outcome to evaluate whether a standard supervised learning risk model is effective in treatment prioritization for MDD adverse events.

Semi-synthetic HTE analysis

In order to evaluate model ability to estimate true HTE we simulate semi-synthetic outcomes utilizing the positive control synthesis approach and selected negative controls based on the methodology outlined in Schuemie et al.[60] This included considering factors such as drug product labels, reports from the FDA on adverse events, relevant publications, and the frequency observed in the study population. We select the psychoactive substance abuse negative control from Chapter 2 as a baseline outcome for semi-synthetic heterogeneity generation.

The negative controls were then used to synthesize the positive controls from a heterogeneous data generating function $h(X) \times T$. We include second order interaction terms in the heterogeneity function to simulate relationships between features.

$$h(X) = \sum_{i=1}^{P} \beta_i x_i + \sum_{j \neq k} \beta_{j,k} x_j x_k$$

The weight parameters (β_i) were drawn from a normal distribution where the number of non-zero weights was varied to simulate different sparsity scenarios. In order to preserve measured confounding[66] between the negative control outcomes (y_{nc}) and model covariates (X), we trained a logistic regression model. This model provided a

predicted probability for each evaluation period that is as a baseline sample probability, prior to the addition of heterogeneity. The probabilities ($\Pr(y_{nc}|X = x)$) are a measure the relationship between the covariates and negative control.

Sampling weights were then calculated by taking the inverse logit of the heterogeneity function plus the confounding function with the addition of an error term $w_{samp} = expit(h(X) + logit(\Pr(y_{nc}|X = x)) + \epsilon)$. Synthetic outcomes were then sampled from a Bernoulli distribution using the sample weights $y_{synth} = Bern(w_{samp})$. We sample the error term from a normal distribution $\epsilon = N(0, \sigma^2)$ and vary σ to simulate additional variance in the data generating function.

Model evaluation

Fitting the models to a matched data set required alteration to standard train-test splitting technique as performing matching prior to the train-test split could potentially result in information leakage from train to test data. The full dataset meeting entry criteria were split into a training and test set at a 1:1 ratio in order balance in the number of treated observations in the test set. If too few treated outcome cases are in the test set performance may be misleading and random variations could result in a split with very few outcomes in the treated group. Logistic regression propensity score matching was then done separately on the training and test sets. In the training set we perform 6:1 matching increase power and model ability to detect heterogeneity. In the test set we perform 2:1 matching resulting in a 7:3 ratio in the final train and test sets.

To train the Causal Forest and Gradient Boosting DML models the hyperparameters number of trees, minimum samples per leaf, minimum samples split, max tree depth were tuned. In addition, an early stopping criteria was also tuned in the Gradient Boosting model, as the Gradient Boosting DML has been shown to be a more flexible estimator and at increased risk of overfitting.[90] Both Gradient Boosting DML and Causal Forest were tuned to optimize rscore,[95] which minimizes deviance between CATE and the dependent variable residuals. Linear DML models used a linear heterogeneity function that was not tuned.

Evaluation of candidate HTE models involved examining a semi-synthetic outcome using metrics such as mean absolute error (MAE) and 95% confidence interval (CI) coverage of synthetic positive controls and Area Under the Targeting Operating Characteristic (AUTOC). AUTOC is a metric used to evaluate treatment prioritization rules. It measures risk reduction of adverse events by not treating patients above a certain risk threshold. This is done by ordering the patients by their estimated CATT, then calculating the risk difference as for increasing treatment risk quantiles. A well-calibrated model will have and increasing trend as the from the risk difference above the threshold to the full population risk. The area under this curve is then calculated and weighted by quantile. AUTOC is calculated directionally, with separate calculations for increased risk and protective AUTOCs. In an increased risk prioritization analysis, a higher AUTOC is better, and vice versa for a protective risk. An AUTOC below zero for an increased risk prioritization analysis indicates the model's risk prioritization is worse than using the population average—in this case the ATT. The AUTOC was calculated on the risk difference scale, so even small values can be meaningful. P-values were calculated by bootstrapping on the test set to provide additional context.

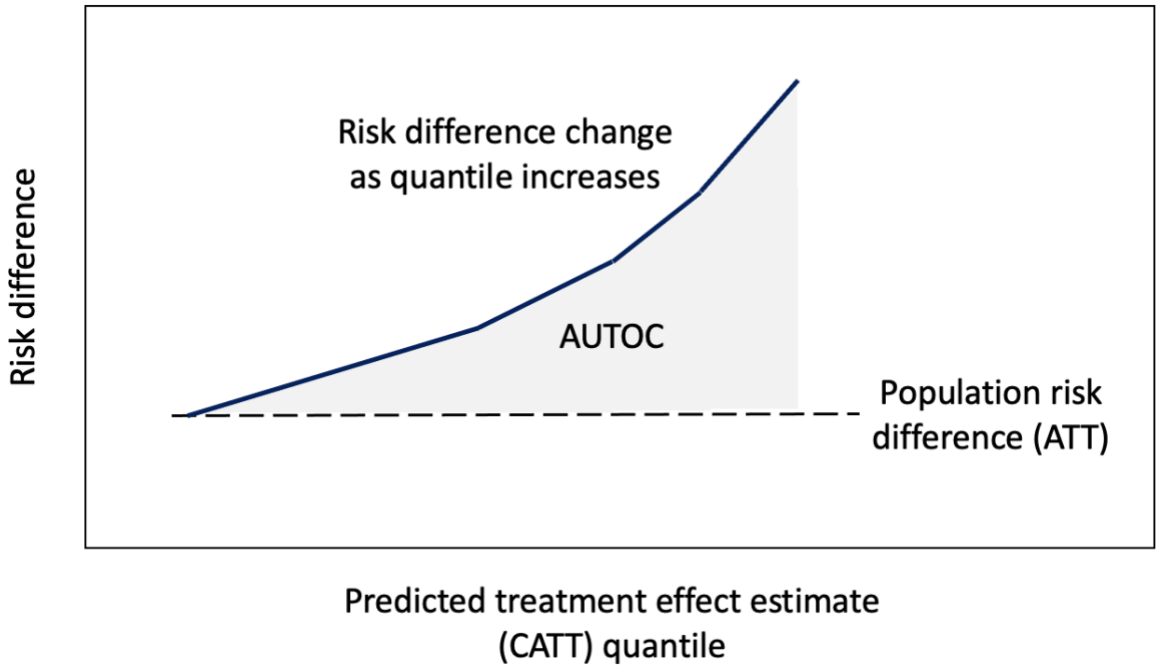


Figure 3-1 AUTOC metric description.

AUTOC is a measure of HTE model calibration. Treatment priority quantile is plotted on the x-axis and treatment risk difference at or greater than the corresponding treatment priority quantile is plotted on the y-axis. The dotted line indicates the population risk difference, because we are working with a covariate matched dataset this is the

ATT. The blue line is the treatment operator curve and displays how the risk difference changes as the treatment priority quantile is increased. An increasing line in this trend indicates good calibration. The area under this curve is the AUTOC. An AUTOC of zero indicates performance no different than using the population average treatment effect.

Evaluation of model calibration on adverse event heterogeneity, including insomnia and nausea, was done by measuring AUTOC, as MAE and CI coverage cannot be calculated without known CATT. To benchmark the evaluation, random forest models were fit to determine risk modeling effectiveness in treatment risk calibration. The random forest models were tuned by grid search cross validation.

Results

Semi-synthetic outcomes

The HTE models were tested under varying conditions of heterogeneity, specifically varying noise and sparsity of feature space. Linear DML demonstrated consistent performance in 95% CI capture rate. Its 95% CI capture was stable as noise increased, in the scenario with 10 non-zero weights the capture rate was 92.6% with 0.0 noise and was 91.2% with noise parameter set to 0.05. In contrast, Causal Forest CI capture decreased as noise increased. In the 10 non-zero weight scenario CI capture was 85.6% with no noise and decreased to 2.5% when the noise parameter was set to 0.05. Causal Forests and Gradient Boosting DML had lowest MAE across heterogeneity scenarios. The MAE increased with noise in each model. Linear DML exhibited higher MAEs compared to Gradient Boosting DML and Causal Forest. AUTOC metric performance varies across models and data generating processes. for the psychoactive substance abuse negative control show in figure 3.2. In the appendix Figure A.13 we include AUTOC performance for the Vitamin D deficiency negative control outcome, which is the negative control with largest number of treated cases.

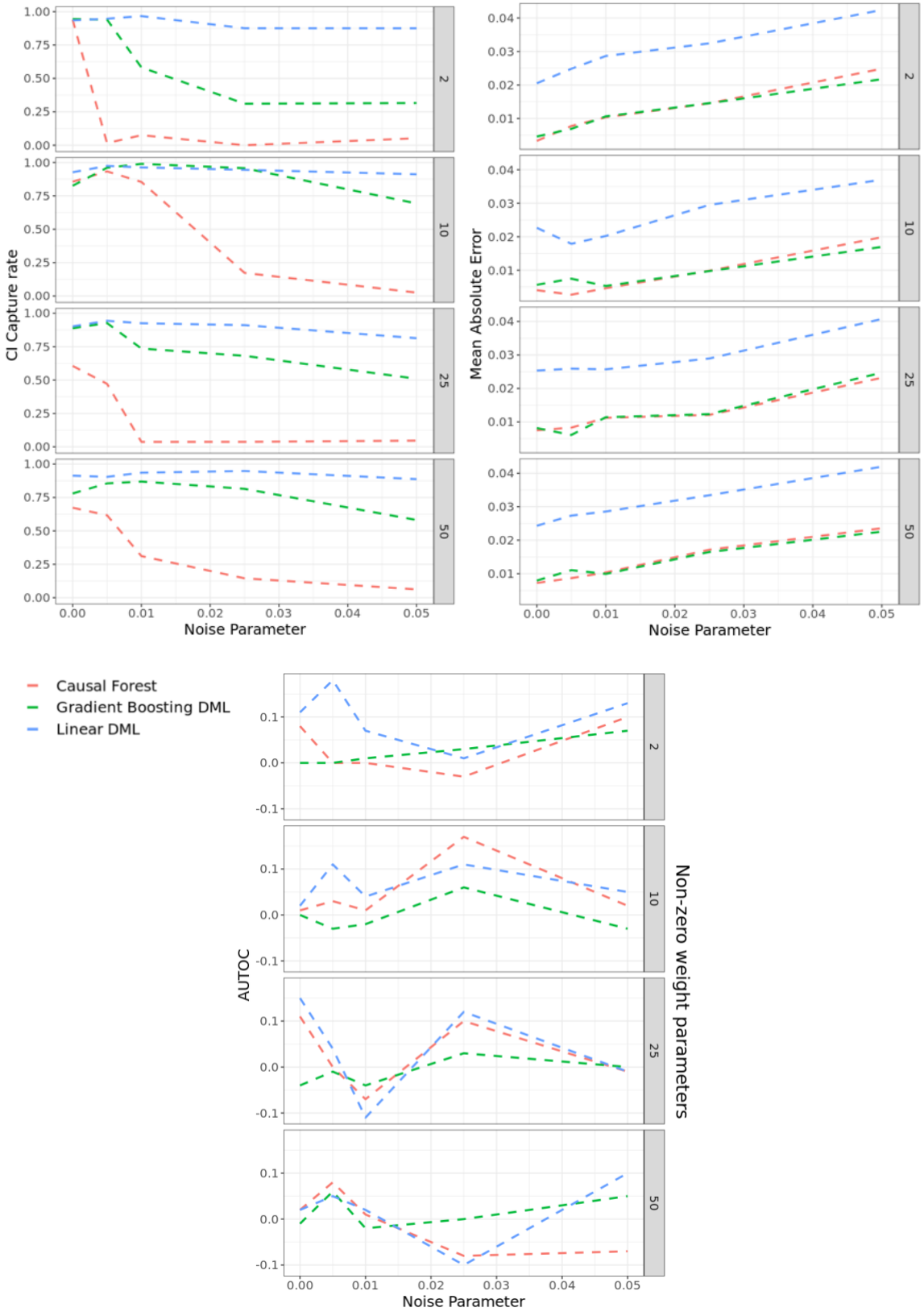


Figure 3-2 Modeling results for semi-synthetic outcomes.

For each outcome model performance trends are plotted as the noise parameter increases. A separate plot is included for each level of the non-zero weight parameter value.

Adverse event outcomes

The train dataset for insomnia included 10,717 evaluation periods, the test dataset included 4,593 evaluation periods. The nausea train set included 12,201 evaluation periods in the train and 5,163 in the test. The insomnia data train set had three variables with absolute standardized mean difference (ASMD) greater than 0.1. ASMD is a measure of covariate balance in a matched dataset, where low ASMD indicates balanced matching.[96] The nausea dataset had four variables with ASMD greater than 0.1.

AUTOC was calculated separately in the increased risk and protective directions. For the insomnia outcome, the Causal Forest model demonstrated the best AUTOC in both protective and increased risk categories, but was not statistically significant in either direction. The training AUTOC for the Causal Forest model in the increased risk direction was 0.96, the test AUTOC reduced by 85%. The Linear DML and Gradient Boosting DML had 89% and 94% reductions in AUTOC from train to test. In the case of nausea, no significant evidence for heterogeneity was identified. Random forest models were included as a benchmark to assess whether risk modeling is sufficient to calibrate risk. For the insomnia outcome, the Random Forest model had Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.62 and Area Under the Precision Recall Curve (AUPRC) of 0.033. For nausea, the model had AUROC of 0.69 and AUPRC of 0.054.

Table 3-1 HTE model calibration on adverse event outcomes

		Increased Risk		Protective	
	Model	AUTOC	p-value	AUTOC	p-value
Insomnia	Causal Forest	**0.14**	**0.13**	**-0.12**	**0.16**
	GB DML	0.03	0.41	0.03	0.59
	Linear DML	0.07	0.30	-0.03	0.41
	Random Forest	-0.06	0.65	0.09	0.77

Nausea	Causal Forest	-0.06	0.82	0.03	0.58
	GB DML	-0.02	0.58	0.03	0.59
	Linear DML	-0.13	0.90	0.17	0.83
	Random Forest	0.08	0.32	-0.03	0.35

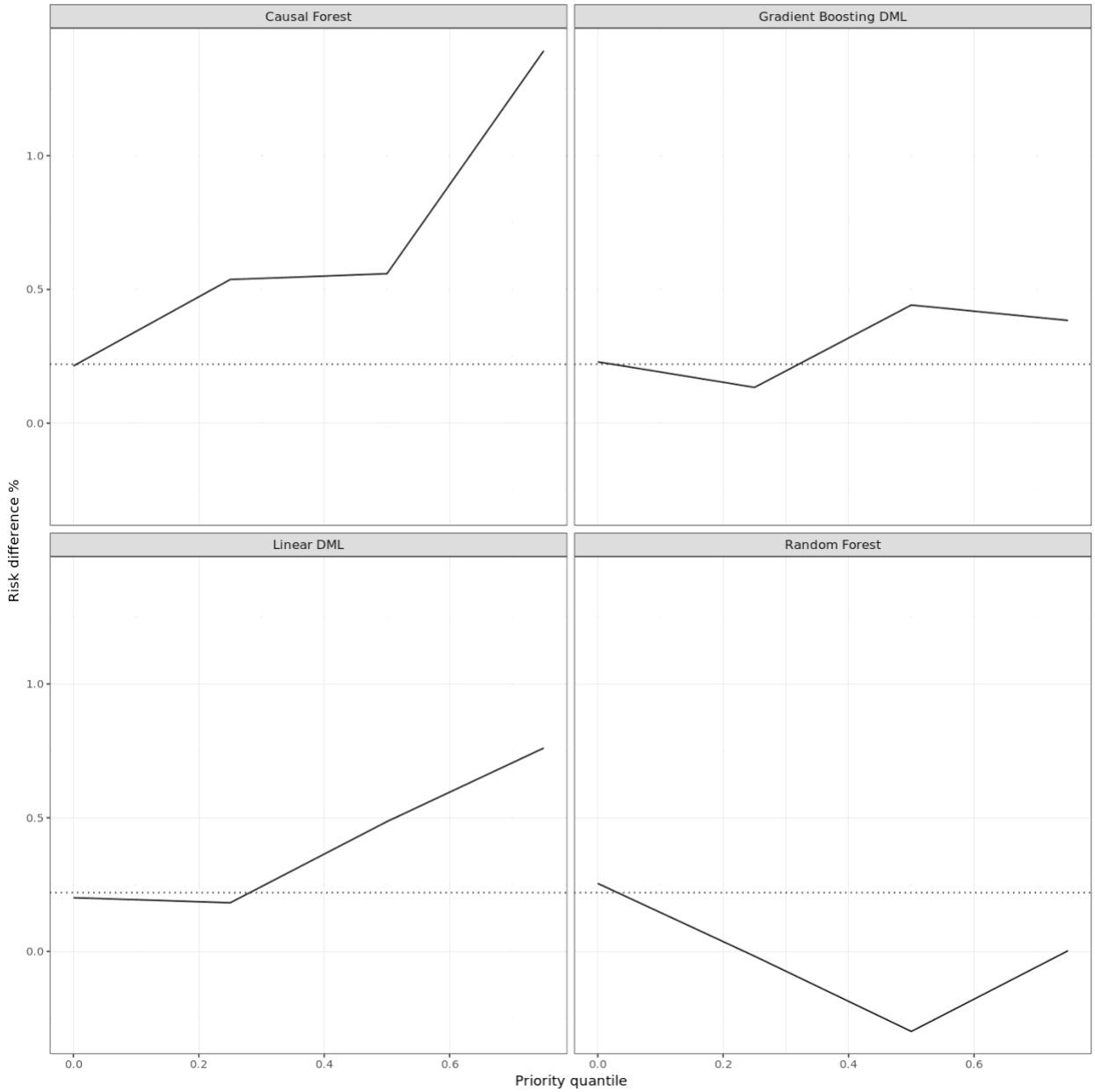


Figure 3-3 Increased risk AUTOC curves for insomnia.

The AUTOC metric measures a HTE models ability to prioritize treatment risk. Treatment priority quantile is plotted on the x-axis and treatment risk difference at or greater than the corresponding treatment priority quantile is plotted on the y-axis. As predicted treatment risk increases, we would expect the observed treatment risk to increase as well. In the increased risk case, an increasing trend is risk difference for higher treatment priority quantiles indicates a well calibrated the model.

Discussion

This study evaluated the performance of HTE models under varying conditions with semi-synthetic outcomes. Overall, as noise was added, each model's performance decreased in MAE and CI capture. Gradient Boosting DML and Causal Forest models had similar MAE. However, Linear DML models had the most stable CI coverage across data generating functions. Causal Forest CI capture deteriorated most significantly as noise was added. Varying feature sparsity did not yield any identifiable difference in model performance. Despite moderate performance in predicting adverse risk, supervised learning-based risk models were not effective in prioritizing treatment of adverse event risk.

We did not find statistically significant evidence for heterogeneity in either of the adverse event outcomes. The Causal Forest had the best AUTOC in both directions for the insomnia outcome. The Causal Forest model also had the least decrease in AUTOC from training to the test set in the insomnia outcome, suggesting less overfitting is taking place. Both the Causal Forest and Gradient Boosting DML followed similar tuning strategies with the exception that an early stopping criterion was added for Gradient Boosting DML. The train AUTOCs in the insomnia outcome were lower for Gradient Boosting DML, it may be that the tuning criteria used led to underfitting of the Gradient Boosting DML.

Heterogeneity in nausea was not detected, and HTE model performance was worse than using the average in each model considered. The underlying reasons for this observed difference in heterogeneity between insomnia and nausea are unclear and warrant further investigation. This could involve exploring the influence of other factors related to SSRI/SNRI prescribing not considered in this study. Prior studies have found a higher prevalence of

nausea than reported in this study,[72] suggesting nausea may be under-ascertained. It is possible that the ascertained nausea cases have many common characteristics that mask heterogeneity.

The study aimed to identify which model is best calibrated to estimate heterogeneity in antidepressant adverse events. In the case of insomnia, calibration was best identified by the Causal Forest model, but the result was not statistically significant. Taking into consideration the results on semi-synthetic data, as well as the adverse event outcome each model had benefits and weaknesses depending on the modeling scenario. The Causal Forest and Gradient Boosting DML reduced MAE relative to Linear DML, but each was ineffective in CI capture as the noise parameter increased, while the Linear DML was consistent. It is unclear at this point, which if any of these metrics would be most meaningful to providers making antidepressant prescribing decisions. It is possible, providers would benefit from knowing which of their patients are at highest risk for an adverse event. Accurate and individualized point estimates and confidence intervals of adverse event risk also have potential to aid providers.

This study is subject to several limitations. Firstly, due to power and data availability, a limited number of outcomes were considered. Future research that models heterogeneity in efficacy with a broader range of clinically meaningful outcomes would be beneficial. Our study indicated that HTE models overfit to training data. The model tuning process requires careful consideration to minimize bias and limited work has been done on the hyper-parameter tuning of HTE models. The tuning objectives can have a substantial impact on the results, suggesting a need for further exploration in this area. There are many additional HTE models worth consideration, including meta-learners that are algorithm agnostic, but require extensive tuning and are higher variance.[97] The results of meta-learners are mixed compared to models included in our study, and tuning and evaluation best practices remain an open area of research.[90,92,98,99]

Conclusion

This study found that HTE models have potential to prioritize insomnia risk when prescribing a new antidepressant. The Linear DML model was best able to calibrate insomnia risk. Heterogeneity was not identified in the nausea outcome. In the semi-synthetic outcomes analysis the Gradient Boosting DML model had best CI capture and the Causal Forest performed best in MAE. Modeling objectives should be carefully considered when choosing metrics

to evaluate performance. Linear DML models assume linear heterogeneity and are less prone to overfit relative to other models considered. Despite moderate performance in outcome risk prediction, supervised learning-based risk models were found to be ineffective in prioritizing treatment for adverse event risk. The low CI capture rates in the causal forest model may be due to bias resulting from hyperparameter tuning, further research on hyper-parameter turning best practices could improve results.

Summary

In this work, we began by developing methods for enhancing prognostic prediction in MDD. We then demonstrated techniques for identification and correction of bias in adverse event effect estimation in observational data. Additionally, we investigated the performance of machine learning models in predicting heterogeneous treatment effects under varying data generating processes and present necessary steps to clinically useful modeling of adverse event heterogeneity.

Chapter 1 of this work focused on the potential benefits of pretraining in improving predictive accuracy of MDD risk models. Our findings suggest that predictive information learned by model weights may be lost at encoding, and that the use of LSTM pretrained models can enhance predictive performance and outperform state-of-the-art predictors in the MDD phenotype. As such, the use of pretrained weights and LSTM architectures may prove useful for future researchers developing risk models in MDD. We also demonstrate that pretrained models can improve predictive accuracy even when trained on data from a single site, which may be of particular interest to researchers and data scientists with limited resources for pretraining foundational EHR neural network models.

Adverse event risk is an important consideration when a new antidepressant is prescribed.[13,14] Electronic health record data capture routine care and are a potentially effective source for post-market safety surveillance. The observational nature of these data can result in biased analyses. In Chapter 2 we utilized empirical calibration to model and correct systematic biases. We observed a protective bias in our ATE estimates, which may be indicative of a healthy user bias or other forms of bias such as differential loss to follow-up. However, our estimates of ATT showed little bias, suggesting that a selection bias may exist in the full study population that decreases when we perform matching.

Heterogeneity of antidepressant adverse events in MDD subpopulations is an important consideration in treatment decisions and patients with higher comorbid burden have been shown to be less likely to be prescribed antidepressants.[16,17] In Chapter 3 we found that HTE models exhibited varied performance depending on the evaluation metric. Causal Forest and Gradient Boosting DML models performed relatively well at error

minimization, but performed worse in CI capture of the true effect. In contrast, Linear DML models displayed consistent performance in CI capture across parameter settings. Performance varied in the AUTOC metric across models and parameter settings and no definite conclusions can be extracted from the results. The AUTOC metric for prioritization has limitations, as this metric calculates risk differences for subsets of the population. It is possible that biases may vary across these subsets. There are other metrics such as the R-score, which has been shown to be effective in HTE model tuning and was utilized for tuning in this study.[95] However, this metric is derived from a loss function and clinically contextualizing may be difficult.[100]

This work has several limitations that should be acknowledged. We have discussed the observational nature of the study and the potential biases that may result. The study was conducted at a single site, which may limit the generalizability of the findings. Adverse event outcomes may have been underreported, and prior antidepressant therapy may have increased the tolerability of antidepressants in the patients studied in Chapters 2 and 3, potentially reducing the overall risk of adverse events in our study population. While the use of HTE models in informatics practice shows promise, the current state of research is still in its infancy, and the lack of understanding of HTE potential impact on clinical practice is a significant limitation.[101] Additionally, most research on HTE models has been conducted using synthetically generated data, there remain research gaps that must be filled before HTE models can aid clinicians.

This work presents many opportunities for future research. Researchers should consider comparing phenotype-specific model pre-training—similar to models from Chapter 1—with foundational full electronic health record (EHR) pre-trained models.[26] Additionally, it may be beneficial to augment in-place prediction models with pre-trained weights, RBM embeddings also showed promise. This approach could serve as an alternative to using large foundational neural network models that may require extensive computational resources.

The work done in chapters 2 and 3 may provide additional insights if the study were conducted at antidepressant initiation, as patients may have increased heterogeneity at this point resulting in more opportunity for impact of HTEs. Studying adverse event risk at initiation may result in increased bias, but patients with prior antidepressant therapy may have increased tolerability of medications reducing heterogeneity in outcomes.[58] Changes in bias

resulting from the change in study population can be examined using the empirical calibration framework. Conducting a multi-site study of these methods would allow for external validation of our results which would increase impact and potentially lead to additional insights. Larger datasets from a multi-site study would allow for analysis of additional treatments and outcomes due to increased study power. Specifically, it would allow us to apply the methods from chapters 2 and 3 to both comparative safety between medications and include additional outcomes such as suicidality.

In order for HTE models to have clinical impact more work is required to understand how evaluation metrics might impact clinical practice. This would likely be best accomplished through qualitative studies that include HTE model training materials, provider interviews, and potentially observation. To our knowledge there is no existing literature on qualitative evaluation of HTEs. However, qualitative studies examining utility of diagnostic machine learning models provide guidance for our work. Pumplun et al[102] examine the adoption process of diagnostic models among clinicians. Their study resulted in categorization of factors that influence adoption of a diagnostic ML tool and provide a framework for assessing maturity of a machine learning adoption process. Sandhu et al[103] show that unfamiliarity with machine learning in the clinical workforce causes concern around perceived utility and trust in the models. This concern may be exacerbated when the technology could be perceived as directive.[104]

Tuning strategies for HTEs need to be refined to overcome the problem of overfitting described in Chapter 3. One of the key issues in this regard is to ensure that the tuning and model fitting objectives are appropriate for each sub-model within the HTE. For example, in propensity score models, the objective is to achieve covariate balance rather than prediction performance.[63] However, the precedent set of using machine learning techniques may lead to application of discrimination models where other forms of optimization may be superior.

Empirical calibration strategies for HTE models may be able to address some of the observed performance issues in chapter 3. The Empirical Calibration framework as it currently stands assumes a linear error model, and more flexible models may be necessary. Additionally, there is limited work on HTE modeling in a matched patient population. Evaluation and implementation is complicated by taking this approach, but recent work using conformal

analysis with HTE models may provide a solution to effectively validating and generating predictions in production.[99]

There is a notable gap in the literature regarding the implementation of useful HTE models. There is a need to identify the key features that make an HTE model effective in practice. Sociotechnical factors must be considered when implementing HTE models, especially with regards to concerns around physician autonomy. Ethical considerations must also be considered when using HTE models. For example, these models estimate treatment effects in subpopulations, which can introduce biases that negatively impact certain groups. It is important to carefully consider these ethical implications before implementing HTE models. Future studies should also investigate issues of fairness in HTE models to ensure that they do not perpetuate existing health disparities.

References

1. Park LT, Zarate CA. Depression in the Primary Care Setting. *N Engl J Med*. 2019;380(6):559-568. doi:10.1056/NEJMcp1712493

2. Ehlman DC. Changes in Suicide Rates — United States, 2019 and 2020. *MMWR Morb Mortal Wkly Rep*. 2022;71. doi:10.15585/mmwr.mm7108a5

3. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5™, 5th Ed*. American Psychiatric Publishing, Inc.; 2013:xliv, 947. doi:10.1176/appi.books.9780890425596

4. Taylor WD. Depression in the Elderly. *N Engl J Med*. 2014;371(13):1228-1236. doi:10.1056/NEJMcp1402180

5. Weissman MM, Bland RC, Canino GJ, et al. Cross-national epidemiology of major depression and bipolar disorder. *JAMA*. 1996;276(4):293-299.

6. Hieronymus F, Emilsson JF, Nilsson S, Eriksson E. Consistent superiority of selective serotonin reuptake inhibitors over placebo in reducing depressed mood in patients with major depression. *Mol Psychiatry*. 2016;21(4):523-530. doi:10.1038/mp.2015.53

7. Mark TL, Levit KR, Buck JA. Datapoints: Psychotropic Drug Prescriptions by Medical Specialty. *Psychiatr Serv*. 2009;60(9):1167-1167. doi:10.1176/ps.2009.60.9.1167

8. Ross EL, Zuromski KL, Reis BY, Nock MK, Kessler RC, Smoller JW. Accuracy Requirements for Cost-effective Suicide Risk Prediction Among Primary Care Patients in the US. *JAMA Psychiatry*. 2021;78(6):642-650. doi:10.1001/jamapsychiatry.2021.0089

9. Si Y, Du J, Li Z, et al. Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. *J Biomed Inform*. 2021;115:103671. doi:10.1016/j.jbi.2020.103671

10. Bagattini F, Karlsson I, Rebane J, Papapetrou P. A classification framework for exploiting sparse multi-variate temporal features with application to adverse drug event detection in medical records. *BMC Med Inform Decis Mak*. 2019;19(1):7. doi:10.1186/s12911-018-0717-4

11. Lasko TA, Denny JC, Levy MA. Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. *PLOS ONE*. 2013;8(6):e66341. doi:10.1371/journal.pone.0066341

12. Tran T, Nguyen TD, Phung D, Venkatesh S. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). *J Biomed Inform*. 2015;54:96-105. doi:10.1016/j.jbi.2015.01.012

13. Thase ME, Friedman ES, Biggs MM, et al. Cognitive therapy versus medication in augmentation and switch strategies as second-step treatments: a STAR*D report. *Am J Psychiatry*. 2007;164(5):739-752. doi:10.1176/ajp.2007.164.5.739

14. Lenze EJ, Mulsant BH, Roose SP, et al. Antidepressant Augmentation versus Switch in Treatment-Resistant Geriatric Depression. *N Engl J Med*. 2023;388(12):1067-1079. doi:10.1056/NEJMoa2204462

15. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Stat Med*. 2014;33(2):209-218. doi:10.1002/sim.5925

16. Gill JM, Klinkman MS, Chen YX. Antidepressant Medication Use for Primary Care Patients with and without Medical Comorbidities: A National Electronic Health Record (EHR) Network Study. *J Am Board Fam Med*. 2010;23(4):499-508. doi:10.3122/jabfm.2010.04.090299

17. Köhler-Forsberg O, Stiglbauer V, Brasanac J, et al. Efficacy and Safety of Antidepressants in Patients With Comorbid Depression and Medical Diseases: An Umbrella Systematic Review and Meta-Analysis. *JAMA Psychiatry*. Published online September 6, 2023. doi:10.1001/jamapsychiatry.2023.2983

18. Vita G, Compri B, Matcham F, Barbui C, Ostuzzi G. Antidepressants for the treatment of depression in people with cancer. *Cochrane Database Syst Rev*. 2023;3(3):CD011006. doi:10.1002/14651858.CD011006.pub4

19. Agabio R, Trogu E, Pani PP. Antidepressants for the treatment of people with co-occurring depression and alcohol dependence. *Cochrane Database Syst Rev*. 2018;4:CD008581. doi:10.1002/14651858.CD008581.pub2

20. Doersch C. Tutorial on Variational Autoencoders. *ArXiv160605908 Cs Stat*. Published online August 13, 2016. Accessed April 30, 2020. http://arxiv.org/abs/1606.05908

21. Baytas IM, Xiao C, Zhang X, Wang F, Jain AK, Zhou J. Patient Subtyping via Time-Aware LSTM Networks. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17. Association for Computing Machinery; 2017:65-74. doi:10.1145/3097983.3097997

22. Jones B, Walsh CG. Unsupervised characterization of Major Depressive Disorder medication treatment pathways. *AMIA Annu Symp Proc*. 2022;2021:591-600.

23. Sagheer A, Kotb M. Unsupervised Pre-training of a Deep LSTM-based Stacked Autoencoder for Multivariate Time Series Forecasting Problems. *Sci Rep*. 2019;9:19038. doi:10.1038/s41598-019-55320-6

24. Zhu S, Zheng W, Pang H. CPAE: Contrastive predictive autoencoder for unsupervised pre-training in health status prediction. *Comput Methods Programs Biomed*. 2023;234:107484. doi:10.1016/j.cmpb.2023.107484

25. Sun Z, Sun Z, Dong W, Shi J, Huang Z. Towards Predictive Analysis on Disease Progression: A Variational Hawkes Process Model. *IEEE J Biomed Health Inform*. 2021;25(11):4195-4206. doi:10.1109/JBHI.2021.3101113

26. Li Y, Rao S, Solares JRA, et al. BEHRT: Transformer for Electronic Health Records. *Sci Rep*. 2020;10(1):7155. doi:10.1038/s41598-020-62922-y

27. Barak-Corren Y, Castro VM, Javitt S, et al. Predicting Suicidal Behavior From Longitudinal Electronic Health Records. *Am J Psychiatry*. 2016;174(2):154-162. doi:10.1176/appi.ajp.2016.16010077

28. Walsh CG, Ribeiro JD, Franklin JC. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *J Child Psychol Psychiatry*. 2018;59(12):1261-1270. doi:10.1111/jcpp.12916

29. Belsher BE, Smolenski DJ, Pruitt LD, et al. Prediction Models for Suicide Attempts and Deaths: A Systematic Review and Simulation. *JAMA Psychiatry*. 2019;76(6):642-651. doi:10.1001/jamapsychiatry.2019.0174

30. Burke TA, Ammerman BA, Jacobucci R. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. *J Affect Disord*. 2019;245:869-884. doi:10.1016/j.jad.2018.11.073

31. Sajjadian M, Lam RW, Milev R, et al. Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis. *Psychol Med*. 2021;51(16):2742-2751. doi:10.1017/S0033291721003871

32. Gao S, Calhoun VD, Sui J. Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neurosci Ther*. 2018;24(11):1037-1052. doi:10.1111/cns.13048

33. Walsh CG, Johnson KB, Ripperger M, et al. Prospective Validation of an Electronic Health Record–Based, Real-Time Suicide Risk Model. *JAMA Netw Open*. 2021;4(3):e211428. doi:10.1001/jamanetworkopen.2021.1428

34. Franklin JC, Ribeiro JD, Fox KR, et al. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychol Bull*. 2017;143(2):187-232. doi:10.1037/bul0000084

35. Wang Y, Blei DM. The Blessings of Multiple Causes. *J Am Stat Assoc*. 2019;114(528):1574-1596. doi:10.1080/01621459.2019.1686987

36. Zhang L, Wang Y, Ostropolets A, Mulgrave JJ, Blei DM, Hripcsak G. The Medical Deconfounder: Assessing Treatment Effects with Electronic Health Records. In: *Machine Learning for Healthcare Conference*. PMLR; 2019:490-512. Accessed September 1, 2021. https://proceedings.mlr.press/v106/zhang19a.html

37. Ranganath R, Perotte A. Multiple Causal Inference with Latent Confounding. *ArXiv180508273 Cs Stat*. Published online March 1, 2019. Accessed September 30, 2021. http://arxiv.org/abs/1805.08273

38. Sun Y, Möller J, Lundin A, Wong SYS, Yip BHK, Forsell Y. Utilization of psychiatric care and antidepressants among people with different severity of depression: a population-based cohort study in Stockholm, Sweden. *Soc Psychiatry Psychiatr Epidemiol*. 2018;53(6):607-615. doi:10.1007/s00127-018-1515-0

39. Snell C, Fernandes S, Bujoreanu IS, Garcia G. Depression, illness severity, and healthcare utilization in cystic fibrosis. *Pediatr Pulmonol*. 2014;49(12):1177-1181. doi:10.1002/ppul.22990

40. Kornstein SG, Joseph AC, Graves WC, Wallenborn JT. Prenatal Depression Severity and Postpartum Care Utilization in a Medicaid Population. *Womens Health Rep New Rochelle N*. 2020;1(1):468-473. doi:10.1089/whr.2020.0079

41. Danciu I, Cowan JD, Basford M, et al. Secondary use of clinical data: the Vanderbilt approach. *J Biomed Inform*. 2014;52:28-35. doi:10.1016/j.jbi.2014.02.003

42. World Health Organization. *ICD-10 : international statistical classification of diseases and related health problems : tenth revision*. World Health Organization; 2004. Accessed October 3, 2022. https://apps.who.int/iris/handle/10665/42980

43. World Health Organization. *International Classification of Diseases : [9th] Ninth Revision, Basic Tabulation List with Alphabetic Index*. World Health Organization; 1978. Accessed October 3, 2022. https://apps.who.int/iris/handle/10665/39473

44. Clinical Classifications Software (CCS) for ICD-10-PCS (beta version). Accessed March 5, 2021. https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp

45. WHO | 2. Anatomical Therapeutic Chemical (ATC) Classification. WHO. Accessed April 30, 2020. http://www.who.int/medicines/regulation/medicines-safety/toolkit_atc/en/

46. Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation. 2014 Measure Updates and Specifications Report Hospital-Wide All-Cause Unplanned Readmission – Version 3.0. Published online July 2014.

47. Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Comput*. 1997;9(8):1735-1780. doi:10.1162/neco.1997.9.8.1735

48. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. In: *Advances in Neural Information Processing Systems*. Vol 30. Curran Associates, Inc.; 2017. Accessed March 17, 2022. https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

49. Home - Keras Documentation. Accessed April 30, 2020. https://keras.io/

50. Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. In: *2017 International Conference on Engineering and Technology (ICET)*. ; 2017:1-6. doi:10.1109/ICEngTechnol.2017.8308186

51. Kessler RC, Stein MB, Petukhova MV, et al. Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *Mol Psychiatry*. 2017;22(4):544-551. doi:10.1038/mp.2016.110

52. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

53. Zeyer A, Bahar P, Irie K, Schlüter R, Ney H. A Comparison of Transformer and LSTM Encoder Decoder Models for ASR. In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. ; 2019:8-15. doi:10.1109/ASRU46091.2019.9004025

54. Yang X, Bian J, Hogan WR, Wu Y. Clinical concept extraction using transformers. *J Am Med Inform Assoc JAMIA*. 2020;27(12):1935-1942. doi:10.1093/jamia/ocaa189

55. Luoma JB, Martin CE, Pearson JL. Contact With Mental Health and Primary Care Providers Before Suicide: A Review of the Evidence. *Am J Psychiatry*. 2002;159(6):909-916. doi:10.1176/appi.ajp.159.6.909

56. Edgcomb JB, Tseng CH, Pan M, Klomhaus A, Zima B. Detection of Suicidal Behavior and Self-harm Among Children Presenting to Emergency Departments: A Tree-based Classification Approach. *AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci*. 2023;2023:108-117.

57. Kennedy SH, Lam RW, McIntyre RS, et al. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder: Section 3. Pharmacological Treatments. *Can J Psychiatry*. 2016;61(9):540-560. doi:10.1177/0706743716659417

58. Wohlreich MM, Mallinckrodt CH, Watkin JG, et al. Immediate switching of antidepressant therapy: results from a clinical trial of duloxetine. *Ann Clin Psychiatry Off J Am Acad Clin Psychiatr*. 2005;17(4):259-268. doi:10.1080/10401230500296402

59. Tomlin A, Reith D, Dovey S, Tilyard M. Methods for retrospective detection of drug safety signals and adverse events in electronic general practice records. *Drug Saf*. 2012;35(9):733-743. doi:10.1007/BF03261970

60. Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, Suchard MA. Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philos Trans R Soc Math Phys Eng Sci*. 2018;376(2128):20170356. doi:10.1098/rsta.2017.0356

61. Hripcsak G, Ryan PB, Duke JD, et al. Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci U S A*. 2016;113(27):7329-7336. doi:10.1073/pnas.1510502113

62. Commissioner O of the. Real-World Evidence. FDA. Published April 25, 2023. Accessed June 6, 2023. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

63. ROSENBAUM PR, RUBIN DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55. doi:10.1093/biomet/70.1.41

64. Hernán MA, Robins JM. *Causal Inference: What If*.

65. Shi X, Miao W, Tchetgen ET. A Selective Review of Negative Control Methods in Epidemiology. *Curr Epidemiol Rep*. 2020;7(4):190-202. doi:10.1007/s40471-020-00243-4

66. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc Natl Acad Sci*. 2018;115(11):2571-2577. doi:10.1073/pnas.1708282114

67. Hripcsak G, Suchard MA, Shea S, et al. Comparison of Cardiovascular and Safety Outcomes of Chlorthalidone vs Hydrochlorothiazide to Treat Hypertension. *JAMA Intern Med*. 2020;180(4):542-551. doi:10.1001/jamainternmed.2019.7454

68. Hoertel N, Sánchez-Rico M, Vernet R, et al. Association between antidepressant use and reduced risk of intubation or death in hospitalized patients with COVID-19: results from an observational study. *Mol Psychiatry*. 2021;26(9):5199-5212. doi:10.1038/s41380-021-01021-4

69. Levintow SN, Nielson CM, Hernandez RK, et al. Pragmatic considerations for negative control outcome studies to guide non-randomized comparative analyses: A narrative review. *Pharmacoepidemiol Drug Saf*. n/a(n/a). doi:10.1002/pds.5623

70. Htoo PT, Measer G, Orr R, et al. Evaluating Confounding Control in Estimations of Influenza Antiviral Effectiveness in Electronic Health Plan Data. *Am J Epidemiol*. 2022;191(5):908-920. doi:10.1093/aje/kwac020

71. Etievant L, Sampson JN, Gail MH. Increasing efficiency and reducing bias when assessing HPV vaccination efficacy by using nontargeted HPV strains. *Biometrics*. Published online March 28, 2022. doi:10.1111/biom.13663

72. Kelly K, Posternak M, Jonathan EA. Toward achieving optimal response: understanding and managing antidepressant side effects. *Dialogues Clin Neurosci*. 2008;10(4):409-418.

73. Yuan K, Zheng YB, Wang YJ, et al. A systematic review and meta-analysis on prevalence of and risk factors associated with depression, anxiety and insomnia in infectious diseases, including COVID-19: a call to action. *Mol Psychiatry*. 2022;27(8):3214-3222. doi:10.1038/s41380-022-01638-z

74. Fang H, Tu S, Sheng J, Shao A. Depression in sleep disturbance: A review on a bidirectional relationship, mechanisms and treatment. *J Cell Mol Med*. 2019;23(4):2324-2332. doi:10.1111/jcmm.14170

75. Oliva V, Lippi M, Paci R, et al. Gastrointestinal side effects associated with antidepressant treatments in patients with major depressive disorder: A systematic review and meta-analysis. *Prog Neuropsychopharmacol Biol Psychiatry*. 2021;109:110266. doi:10.1016/j.pnpbp.2021.110266

76. Weight Trimming and Propensity Score Weighting | PLOS ONE. Accessed August 10, 2023. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018174

77. Izurieta HS, Chillarige Y, Kelman J, et al. Relative Effectiveness of Influenza Vaccines Among the United States Elderly, 2018–2019. *J Infect Dis*. 2020;222(2):278-287. doi:10.1093/infdis/jiaa080

78. Zeileis A. Econometric Computing with HC and HAC Covariance Matrix Estimators. *J Stat Softw*. 2004;11:1-17. doi:10.18637/jss.v011.i10

79. Nawi AM, Ismail R, Ibrahim F, et al. Risk and protective factors of drug abuse among adolescents: a systematic review. *BMC Public Health*. 2021;21(1):2088. doi:10.1186/s12889-021-11906-2

80. Sullivan MD. Depression Effects on Long-term Prescription Opioid Use, Abuse, and Addiction. *Clin J Pain*. 2018;34(9):878-884. doi:10.1097/AJP.0000000000000603

81. Urits I, Gress K, Charipova K, et al. Cannabis Use and its Association with Psychological Disorders. *Psychopharmacol Bull*. 2020;50(2):56-67.

82. Webster LR. Risk Factors for Opioid-Use Disorder and Overdose. *Anesth Analg*. 2017;125(5):1741-1748. doi:10.1213/ANE.0000000000002496

83. Carvalho AF, Sharma MS, Brunoni AR, Vieta E, Fava GA. The Safety, Tolerability and Risks Associated with the Use of Newer Generation Antidepressant Drugs: A Critical Review of the Literature. *Psychother Psychosom*. 2016;85(5):270-288. doi:10.1159/000447034

84. Mackay FJ, Dunn NR, Wilton LV, Pearce GL, Freemantle SN, Mann RD. A comparison of fluvoxamine, fluoxetine, sertraline and paroxetine examined by observational cohort studies. *Pharmacoepidemiol Drug Saf*. 1997;6(4):235-246. doi:10.1002/(SICI)1099-1557(199707)6:4<235::AID-PDS293>3.0.CO;2-3

85. Shrank WH, Patrick AR, Alan Brookhart M. Healthy User and Related Biases in Observational Studies of Preventive Interventions: A Primer for Physicians. *J Gen Intern Med*. 2011;26(5):546-550. doi:10.1007/s11606-010-1609-1

86. Nelson JC. Treatment of antidepressant nonresponders: augmentation or switch? *J Clin Psychiatry*. 1998;59 Suppl 15:35-41.

87. Segal JB, Varadhan R, Groenwold RHH, et al. Assessing Heterogeneity of Treatment Effect in Real-World Data. *Ann Intern Med*. 2023;176(4):536-544. doi:10.7326/M22-1510

88. Wager S, Athey S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *ArXiv151004342 Math Stat*. Published online July 9, 2017. Accessed May 9, 2022. http://arxiv.org/abs/1510.04342

89. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci*. 2016;113(27):7353-7360. doi:10.1073/pnas.1510489113

90. Powers S, Qian J, Jung K, et al. Some methods for heterogeneous treatment effect estimation in high dimensions. *Stat Med*. 2018;37(11):1767-1787. doi:10.1002/sim.7623

91. Pearl J. An Introduction to Causal Inference. *Int J Biostat*. 2010;6(2). doi:10.2202/1557-4679.1203

92. Curth A, Svensson D, Weatherall J, Schaar M van der. Really Doing Great at Estimating CATE? A Critical Look at ML Benchmarking Practices in Treatment Effect Estimation. In: ; 2021. Accessed September 22, 2023. https://openreview.net/forum?id=FQLzQqGEAH

93. Holland PW. Statistics and Causal Inference. *J Am Stat Assoc*. 1986;81(396):945-960. doi:10.1080/01621459.1986.10478354

94. Wendling T, Jung K, Callahan A, Schuler A, Shah NH, Gallego B. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Stat Med*. 2018;37(23):3309-3324. doi:10.1002/sim.7820

95. Schuler A, Baiocchi M, Tibshirani R, Shah N. A comparison of methods for model selection when estimating individual treatment effects. Published online June 13, 2018. doi:10.48550/arXiv.1804.05146

96. Ho DE, Imai K, King G, Stuart EA. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Polit Anal*. 2007;15(3):199-236. doi:10.1093/pan/mpl013

97. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc Natl Acad Sci*. 2019;116(10):4156-4165. doi:10.1073/pnas.1804597116

98. Shi J, Norgeot B. Learning Causal Effects From Observational Data in Healthcare: A Review and Summary. *Front Med*. 2022;9. Accessed September 22, 2023. https://www.frontiersin.org/articles/10.3389/fmed.2022.864882

99. Alaa A, Ahmad Z, van der Laan M. Conformal Meta-learners for Predictive Inference of Individual Treatment Effects. Published online August 28, 2023. doi:10.48550/arXiv.2308.14895

100. Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*. 2021;108(2):299-319. doi:10.1093/biomet/asaa076

101. Bica I, Alaa AM, Lambert C, van der Schaar M. From Real-World Patient Data to Individualized Treatment Effects Using Machine Learning: Current and Future Methods to Address Underlying Challenges. *Clin Pharmacol Ther*. 2021;109(1):87-100. doi:10.1002/cpt.1907

102. Pumplun L, Fecho M, Wahl N, Peters F, Buxmann P. Adoption of Machine Learning Systems for Medical Diagnostics in Clinics: Qualitative Interview Study. *J Med Internet Res*. 2021;23(10):e29301. doi:10.2196/29301

103. Sandhu S, Lin AL, Brajer N, et al. Integrating a Machine Learning System Into Clinical Workflows: Qualitative Study. *J Med Internet Res*. 2020;22(11):e22421. doi:10.2196/22421

104. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012;98(9):691-698. doi:10.1136/heartjnl-2011-301247

105. Bulus M. pwrss: Statistical Power and Sample Size Calculation Tools. Published online April 11, 2023. Accessed September 11, 2023. https://cran.r-project.org/web/packages/pwrss/index.html

Appendix

Table A.1-Depression ICD codes list

	Code	Description
ICD-9	311.x	Depressive disorder, not elsewhere classified
	296.2x	Major depressive disorder single episode
	296.3x	Major depressive disorder recurrent episode
	300.4x	Dysthymic disorder
ICD-10	F32.xx	Major depressive disorder, single episode
	F33.xx	Major depressive disorder, recurrent
	F34.1	Dysthymic disorder

Table A.2-Antidepressant List with ATC code

Code	Name	Code	Name	Code	Name
N06BA04	methylphenidate	N06AA10	nortriptyline	N06AX16	venlafaxine
N06BA02	dexamfetamine	N06AA11	protriptyline	N06AX17	milnacipran
N05AH04	quetiapine	N06AA04	clomipramine	N06AX06	nefazodone
N05AH03	olanzapine	N06AA21	maprotiline	N06AX23	desvenlafaxine
N05AE04	ziprasidone	N06AA17	amoxapine	N03AX09	lamotrigine
N05AN01	lithium	N06AA02	imipramine	N05AX12	aripiprazole
N06AF04	tranylcypromine	N06AX24	vilazodone	N06AB04	citalopram
N06AF03	phenelzine	N06AX12	bupropion	N06AB03	fluoxetine
N06AF01	isocarboxazid	N06AX21	duloxetine	N06AB05	paroxetine
N06AA12	doxepin	N06AX26	vortioxetine	N06AB10	escitalopram
N06AA09	amitriptyline	N06AX05	trazodone	N06AB08	fluvoxamine
N06AA01	desipramine	N06AX11	mirtazapine	N06AB06	sertraline

Table A.3 Autoencoder model parameters

Attention Encoder	Layers
	Multi-head-attention(head size = 64, number of heads = 4)
	Convolutional 1D(filters = 294, kernel size = 6, activation = 'relu')
	Convolutional 1D(filters = 147, kernel size = 3, activation = 'relu')
Attention Decoder	
	Multi-head-attention(head size = 64, number of heads = 4)
	Multi-head-attention(head size = 64, number of heads = 4)
	Convolutional 1D(filters = 294, kernel size = 6, activation = 'relu')
	Time distributed dense(units = 147, activation = 'sigmoid')
LSTM Encoder	
	LSTM(units=147, activation='relu')

	LSTM(units=147, activation='relu')
LSTM Decoder	
	LSTM(units=147, activation='relu')
	LSTM(units=147, activation='relu')
	Time distributed dense(units = 147, activation = 'sigmoid')

Table A.4-Antidepressant for SSRI and SNRI referenced in Chapters 2 and 3

Clinical Grouping	Medication Name (ATC Level 5)
Selective serotonin reuptake inhibitors (SSRI)	citalopram fluoxetine paroxetine escitalopram fluvoxamine sertraline
Serotonin-norepinephrine reuptake inhibitors (SNRI)	desvenlafaxine duloxetine venlafaxine milnacipran

Table A.5-Adverse event SNOMED CT codes referenced in Chapters 2 and 3

Outcome	**SNOMED CT code**	**Description**
Suicidal Ideation and Behavior	439235	Self-inflicted injury
	4181216	Self-administered poisoning
	444362	Suicidal deliberate poisoning
	4273391	Suicidal thoughts
	440925	Suicide
	4303690	Intentionally harming self
Insomnia	439708	Disorders of initiating and maintaining sleep
	436962	Insomnia
	4305303	Sleep deprivation
Nausea	31967	Nausea
	30284 (exclude)	Motion sickness

Table A.6-Negative control outcome concept IDs referenced in Chapter 2

Concept ID code	**Description**	**Treated case count**
436070	Vitamin D deficiency	84
436230	Blood chemistry abnormal	67
4150062	Knee pain	62
437390	Hypoxemia	44
437677	Abnormal findings on diagnostic imaging of lung	41

77646	Disorder of bone and articular cartilage	40
434004	Hypervolemia	28
200528	Ascites	25
132736	Bacteremia	25
4110815	Sensorineural hearing loss bilateral	24
140648	Onychomycosis due to dermatophyte	23
314754	Wheezing	23
443257	Swelling-lump finding	22
440276	Infection AND/OR inflammatory reaction due to internal prosthetic device implant AND/OR graft	21
136788	Spinal stenosis of lumbar region	21

Table A.7-Negative control outcomes grouped, referenced in Chapter 2.

ancestor concept ID code	Ancestor description	Treated case count	Grouped concepts	Number of concepts grouped
433125	Infection due to Staphylococcus aureus	31	Infection by methicillin sensitive Staphylococcus aureus 40481816, Methicillin resistant Staphylococcus aureus infection 440940	2
4252534	Disease due to Gram-negative bacillus	24	Bacterial infection due to Pseudomonas 438064, Infection due to Escherichia coli 440320	2
4239381	Psychoactive substance abuse	21	Cannabis abuse 434327, Opioid abuse 438130	2
443238	Diabetic - poor control	21	Type II diabetes mellitus uncontrolled 40482801, Type 1 diabetes mellitus uncontrolled 40484648	2
4018050	Localized infection	20	Posttraumatic wound infection 4153877, Localized infection of skin ANDOR subcutaneous tissue 443600	2
4161193	Disease due to Gram-positive bacteria	20	Septicemia due to enterococcus 133956, Staphylococcal infectious disease 435459	2
4090739	Nutritional disorder	31	Moderate protein calorie malnutrition weight for age 6074 percent of standard 4098458, Severe protein calorie malnutrition Gomez less than 60 percent of standard weight 4233565, Malnutrition of moderate degree Gomez 60 percent to less than 75 percent of standard weight 436078	3
254068	Disorder of upper respiratory system	28	Hypertrophy of tonsils 28457, Deviated nasal septum 377910, Hypertrophy of nasal turbinates 440129	3
73008	Enthesopathy	21	Enthesopathy of hip region 198846, Enthesopathy of foot region 4347178, Spinal enthesopathy 75347	3

4042141	Ear and auditory finding	20	Impacted cerumen 374375, Hearing loss of right ear 43021778, Asymmetrical sensorineural hearing loss 443577	3
4100932	Knee joint finding	31	Tear of meniscus of knee 4035415, Disorder of patellofemoral joint 4035422, Knee joint effusion 4115991, Current tear of medial cartilage ANDOR meniscus of knee 80242	4
436670	Metabolic disease	21	Disorder of the urea cycle metabolism 434311, Disorders of bilirubin excretion 434887, Lipoprotein deficiency disorder 435516, Disorder of lipid metabolism 437530	4
4302537	Digestive system finding	21	Viral hepatitis without hepatic coma 193693, Diverticulosis of large intestine without diverticulitis 4164898, Ileostomy present 4201717, Chronic viral hepatitis B without delta agent 439674, Viral hepatitis in mother complicating childbirth 45757141	5
4249437	Disease due to Alphaherpesvirinae	20	Disseminated herpes zoster 4205455, Herpes simplex with complication 438962, Herpes simplex without complication 440021, Herpes zoster without complication 440329, Herpes simplex 444429	5
441969	Radiology result abnormal	45	Mammography abnormal 4059049, Abnormal findings on diagnostic imaging of limbs 4171776, Abnormal findings on diagnostic imaging of breast 434169, Abnormal findings diagnostic imaging heart coronary circulat 435081, Abnormal findings on diagnostic imaging of skull and head 439154, Abnormal findings on diagnostic imaging of urinary organs 440529	6
320136	Disorder of respiratory system	26	Traumatic pneumothorax without open wound into thorax 253896, Bronchiectasis 256449, Pleural plaque 4050884, Foreign body in bronchus 443287, Acute exacerbation of asthma 45771045, Acute exacerbation of mild persistent asthma 46270082	6
4042837	Disorder of neck	21	Cervical spondylosis with myelopathy 136198, Cervical spine ankylosis 4001454, Cervical disc disorder with radiculopathy 4067313, Late effect of fracture of cervical vertebra 4194739, Spinal stenosis in cervical region 436785, Displacement of cervical intervertebral disc without myelopathy 74725, Disorder of cervical spine 80497	7
4022449	Finding of shoulder region	20	Disorder of joint of shoulder region 40484571, Nontraumatic rotator cuff tear 4172970, Impingement syndrome of shoulder region 4344500, Derangement of shoulder 45757404, Full thickness rotator cuff tear 73564, Disorder of shoulder 77630, Adhesive capsulitis of shoulder 77644, Articular cartilage disorder of shoulder region 77955	8

4024000	Urinary system finding	20	Incomplete emptying of bladder 193020, Urinary bladder stone 193520, Mechanical complication due to urethral indwelling catheter 194847, Poor stream of urine 4012231, Atrophy of kidney 4058977, Absent kidney 4092879, Lower urinary tract symptoms 443350, Sign or symptom of the urinary system 77673	8
4027384	Inflammatory disorder	24	Acute osteomyelitis of pelvic region and or thigh 133570, Biceps tendinitis 4000968, Systemic inflammatory response syndrome associated with organ dysfunction 40479649, Rheumatic endocarditis 4169568, Systemic inflammatory response syndrome 434821, Contact dermatitis due to plants except food 444375, Tibialis tendinitis 77081, Achilles tendinitis 77963, Tuberculosis of vertebral column 81496	9
435726	Mechanical complication of device	24	Mechanical complication of dialysis catheter 4070976, Displacement of internal fixation device 43022016, Mechanical complication due to coronary bypass graft 432499, Mechanical complication of device 435726, Mechanical complication of cardiac device implant ANDOR graft 438297, Mechanical complication of peritoneal dialysis catheter 440302, Mechanical complication of genitourinary device implant ANDOR graft 442012, Breakage of joint prosthesis 80008, Mechanical complication of internal joint prosthesis 80269, Prosthetic joint loosening 80286	10
43021974	Complication associated with device	21	Disorder of cardiovascular prostheses and implants 142026, Complication associated with insulin pump 43021246, Infection associated with implant 43021258, Complication associated with device 43021974, Disorders of urogenital prostheses or implants 76887	5
442019	Complication of procedure	40	Subcutaneous emphysema resulting from a procedure 138056, Disorders of prostheses and implants of the nervous system 373105, Stenosis due to any device implant ANDOR graft 4008710, Complication of artificial skin graft and decellularized allodermis 4207606, Complication of surgical procedure 434547, Complication of gastrostomy 434675, Late effect of medical and surgical care complication 434814, Foreign body accidentally left during a procedure 442018, Gastric band procedure complication 45757691	9

40484102	Abnormal finding on evaluation procedure	45	Abnormal results of cardiovascular function studies 137989, Imaging of gastrointestinal tract abnormal 40482267, Cerebrospinal fluid examination abnormal 4065770, Urine cytology abnormal 4150384, Abnormal findings on microbiological examination of urine 4168689, Abnormal cytological findings in CSF 4168693, Abnormal results function studies of central nervous system 432451, Abnormal cervical smear 434165, Atypical squamous cells of undetermined significance on cervical Papanicolaou smear 434170, Electromyogram abnormal 441415, Atypical squamous cells of undetermined significance on anal Papanicolaou smear 443709	11
37311678	Finding of abdominopelvic segment of trunk	27	Hypersplenism 192298, Malignant ascites 192735, Pelvic organ injury without open wound into abdominal cavity 195682, Contusion of abdominal wall 196569, Lumbar spine ankylosis 4002140, Splenic infarction 4044745, Lower back injury 4151985, Intraabdominal and pelvic swelling mass and lump 4168222, Closed fracture of lumbar vertebra without spinal cord injury 435933, Genitourinary tract infection in pregnancy not delivered 74415, Disorder of coccyx 78235	11
4339410	Disorder of skeletal system	20	Closed fracture of thoracic vertebra without spinal cord injury 316535, Disorder of intervertebral disc of thoracic spine 321389, Disorder of mastoid 373216, Cervicothoracic ankylosis 4002139, Elbow joint effusion 4117881, Wrist joint effusion 4117883, Collapse of thoracic vertebra 4203555, Chondrocalcinosis 437064, Articular disc disorder of temporomandibular joint 74399, Arthropathy associated with a neurological disorder 74723, Hypertrophic osteoarthropathy 74731, Polyarthropathy 75897, Closed fracture of vertebral column without spinal cord injury 77403, Aseptic necrosis of bone 77650, Effusion of joint of hand 78834	15
43530815	Traumatic injury by site	21	Open wound of forehead without complication 138896, Open wound of face without complication 140259, Abrasion and or friction burn of trunk without infection 199192, Open wound of head without complication 372765, Injury of nose 4024306, Open wound of front wall of thorax 4050089, Abrasion of trunk 4050704, Superficial injury of hand 4086197, Laceration of upper arm 4152933, Laceration of forearm 4155034, Scalp laceration 4166902, Abrasion and or friction burn of multiple sites 443585, Injury of elbow 444189, Open wound of finger with complication 74806, Open wound of hand except fingers with	16

	complication 75686, Open wound of wrist without complication 78593

Covariate Balance and Power Analysis Table A.8

	Covariate Balance		Power	
k:1 matching strategy	**Covariates w/ ASMD > 0.05**	**Covariates w/ ASMD > 0.1**	**Insomnia**	**Nausea**
1:1	1	0	0.40	0.41
2:1	2	0	0.55	0.57
3:1	5	1	0.68	0.69
4:1	7	2	0.77	0.78
5:1	8	3	0.84	0.85
6:1	8	4	0.88	0.90

Evaluation of k:1 matching results are displayed in Table A.6. For each k, we assess covariate balance by the number of features with absolute standardized mean difference greater than 0.05 and 0.1. Power analysis was conducted for each outcome using the pwrss library in R.[1] We evaluated the power of detection of a statistically significant coefficient in a logistic regression model with risk ratio of 1.2, with baseline probabilities calculated for each outcome. The sample size corresponded to the number of samples available under the k:1 matching strategy. We also account for treatment-covariate multicollinearity with an assumption of 20% variation in treatment explained by covariates. Due to the infrequency of the SISB outcome, power does not reach an acceptable magnitude for values of k considered—at k=4 power is 0.22. Therefore, SISB is excluded from the study analysis. Insomnia and Nausea exceed 0.75 power at k=4. Though at k=4, covariate balance has decreased, we make this trade off in order to preserve a reasonable chance of effect detection. For the full dataset power was at least 0.99 for each outcome.

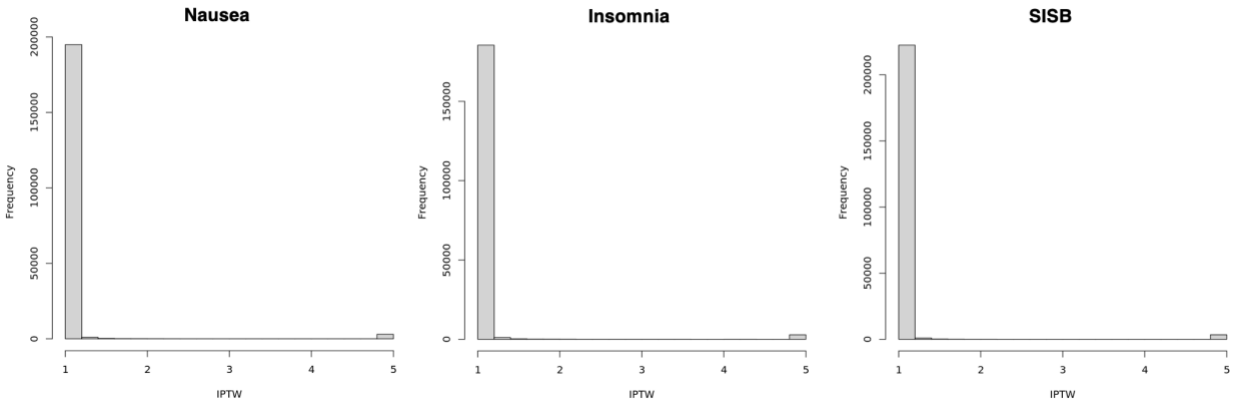
Matching analysis covariate balance by outcome Table A.9

Outcome	**N treated**	**N control**	**Count of covariates w/ ASMD > 0.05**	**Covariates w/ ASMD > 0.05**

1 105

Nausea	3,420	13,680	6	Psychiatry referral, prior ED visit, prior high utilization, alcohol related disorders, substance related disorders, prior SISB
Insomnia	3,144	12,576	5	Psychiatry referral, prior ED visit, alcohol related disorders, substance related disorders, prior SISB

IPTW Distributions by outcome Figure A.10



In figure A.8 we display histograms of the IPTW weights for each outcome. The weights are truncated at 5. The IPTW distributions are similar across outcomes with the majority of IPTW density is near 1. The medians for each IPTW distribution range from 1.011- 1.012.

Figure A.11 follow up visit rate for study inclusion criteria

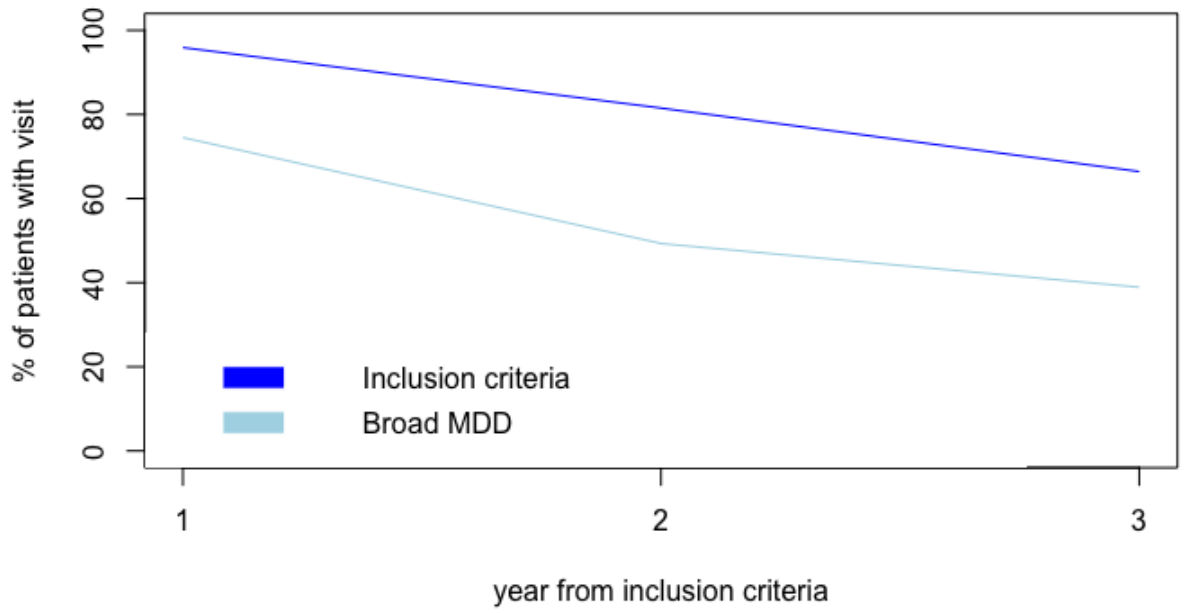
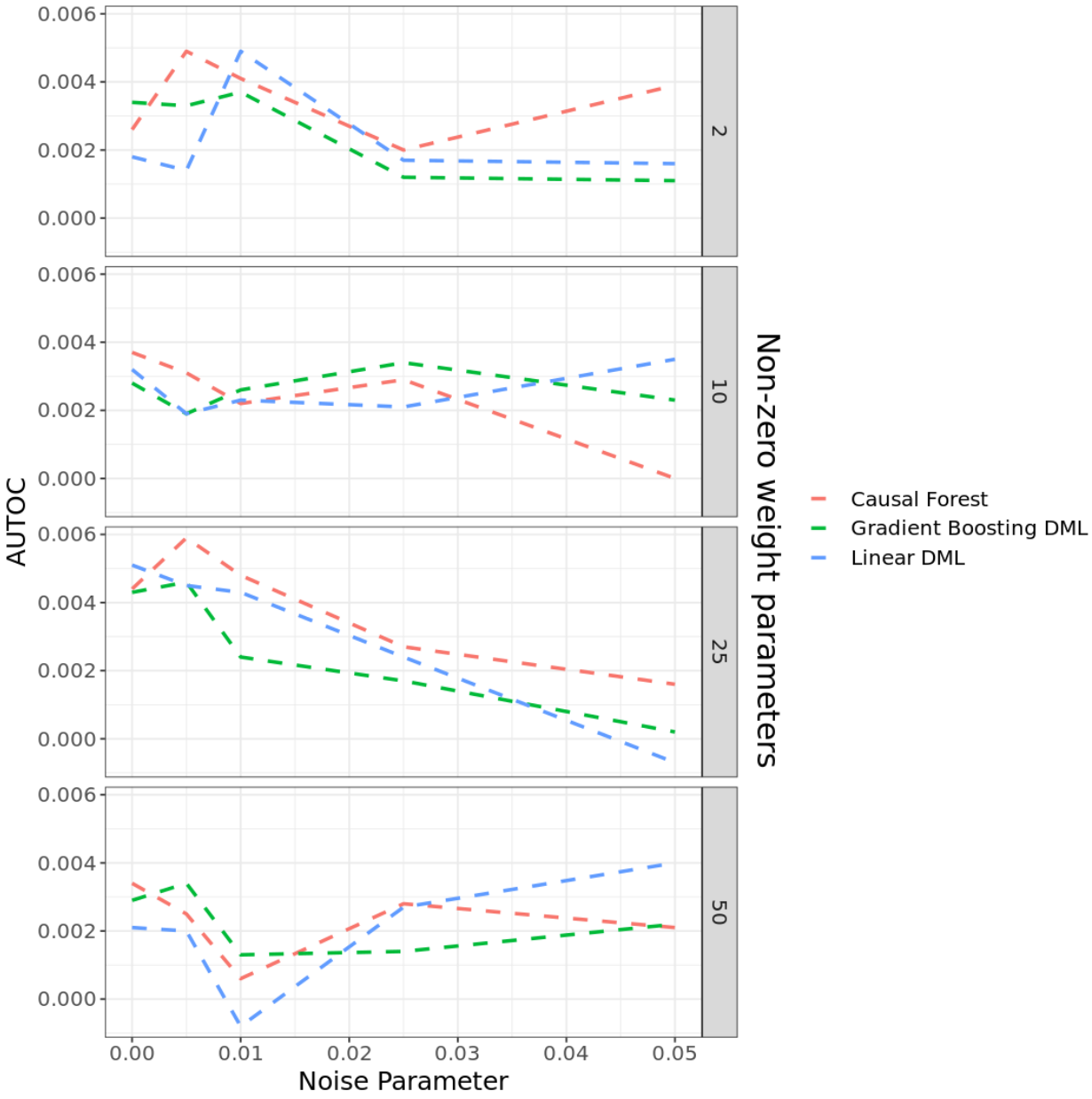


Figure A.9 compares the follow up visit rate of patients with the study inclusion criteria to those with a broader MDD definition at 1, 2, and 3 years following meeting inclusion criteria. The Broad MDD population includes MDD indication, age 18-90, bipolar and schizophrenia excluded.

Table A.12 Bipolar and Schizophrenia Exclusion codes

Diagnosis	Version	ICD Code	Description
Bipolar	ICD9	296.*	Bipolar I disorder, single manic episode, unspecified
		293.81	Psychotic disorder with delusions in conditions classified elsewhere
		293.82	Psychotic disorder with hallucinations in conditions classified elsewhere
	ICD10	F31.*	Bipolar disorder
Schizo-phrenia	ICD9	295.*	Simple type schizophrenia, unspecified
		297.*	Paranoid state, simple
		298.*	Depressive type psychosis
	ICD10	F06.0	Psychotic disorder with hallucinations due to known physiological condition
		F06.2	Psychotic disorder with delusions due to known physiological condition
		F20.0	Paranoid schizophrenia
		F20.1	Disorganized schizophrenia
		F20.2	Catatonic schizophrenia
		F20.3	Undifferentiated schizophrenia
		F20.5	Residual schizophrenia
		F20.81	Schizophreniform disorder
		F20.89	Other schizophrenia
		F20.9	Schizophrenia, unspecified
		F21.*	Schizotypal disorder
		F22.*	Delusional disorders
		F23.*	Brief psychotic disorder
		F24.*	Shared psychotic disorder
		F25.0	Schizoaffective disorder, bipolar type
		F25.1	Schizoaffective disorder, depressive type
		F25.8	Other schizoaffective disorders
		F25.9	Schizoaffective disorder, unspecified
		F28.*	Other psychotic disorder not due to a substance or known physiological condition
		F29.*	Unspecified psychosis not due to a substance or known physiological condition

Figure A.13-AUTOC for semi-synthetic outcomes with the Vitamin D Deficiency negative control outcome


AUTOC
0.006
0.004
0.002
0.000
2
10
25
50
0.00
0.01
0.02
0.03
0.04
0.05
Noise Parameter
Non-zero weight parameters
Causal Forest
Gradient Boosting DML
Linear DML