Using high-dimensional pharmacogenomics data to predict effective antidepressant treatment response and symptom remission in major depressive disorder patients

by

Lauren Michelle Rost

B.A., St. Mary's College of Maryland, 2015

M.S., University of Pittsburgh, 2019

Submitted to the Graduate Faculty of the

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Lauren Michelle Rost

It was defended on

July 13, 2021

and approved by

Philip Empey, Associate Professor, Pharmacy and Therapeutics

Shyam Visweswaran, Associate Professor, Biomedical Informatics

Tanya Fabian, Associate Professor, Pharmacy and Therapeutics

Gerry Douglas, Assistant Professor, Department of Biomedical Informatics

Uma Chandran, Research Associate Professor, Biomedical Informatics

Dissertation Director: Douglas Landsittel, Professor and Chair, Epidemiology and Biostatistics, Indiana University-Bloomington

Using high-dimensional pharmacogenomics data to predict effective antidepressant treatment response and symptom remission in major depressive disorder patients

Lauren Michelle Rost, PhD

University of Pittsburgh, 2021

Background: Major depressive disorder (MDD) is a highly prevalent, chronic and disabling condition. Antidepressants are the mainstay of treatment with selective serotonin reuptake inhibitors (SSRIs) recommended as first-line treatment. However, antidepressant response rates are dismal with only 35-45% of patients achieving remission after initial agent. Patients with MDD are often exposed to a series of antidepressants in a trial-and-error process in effort to achieve symptom remission or treatment response. We hypothesize that utilization of patients' electronic health record (EHR) and machine learning methods can improve MDD treatment outcome prediction.

Methods: Clinical and pharmacy data were extracted from the UPMC EHR and utilized to examine MDD electronic phenotyping in addition to characterizing antidepressant treatment outcomes including dose changes, treatment sequences, and combinations. In addition, EHR features associated with predicting MDD treatment outcomes were explored. A reproducible pipeline was constructed to yield reproducible results with other data sources, including the addition of PGx data.

Results: SSRIs were the most common initial antidepressant class prescribed for MDD patients, followed by SNRIs and NDRIs. The most common initial antidepressant prescribed for patients were SSRIs: sertraline, citalopram, and escitalopram, respectively. Early depression patients, those responding to initial antidepressants, comprised 39.69% of the analysis cohort, while 60.31% of

patients required a medication switch or augmentation. The most commonly prescribed two-drug sequence was citalopram then bupropion. When examining the probabilities of transitioning between antidepressant classes, transition probabilities to SSRIs were the highest. The highest performing machine learning model for predicting treatment response was a random forest using the top 25 clinical features (accuracy: 77.21%, F1-score: 87.07%), while the best model for predicting symptom remission was a generalized linear model using the top 25 features (accuracy: 68.16%, F1-score: 33.33%).

Discussion: SSRIs are commonly prescribed to patients with MDD, not only as first-line treatment, but are just as likely to be revisited throughout the treatment course. Future directions include assessing the value-add of PGx data in predicting antidepressant treatment response, validating results using EHR data from other health systems with more diverse patient populations, and implementing the prediction model in clinical practice to inform antidepressant treatment selection.

Table of Contents

List of Tables

List of Figures

1.0 Background

1.1 Major Depressive Disorder

Major depressive disorder (MDD) is a complex, heterogenous disorder with multiple risk factors and underlying biological mechanisms [1,2]. The Diagnostic and Statistical Manual of Mental Disorders (DSM) has been systematic in symptom-based classification and diagnosis of MDD patients since 1952. An MDD diagnosis often entails a change of mood, sadness, or irritability with a myriad of psychophysiological aspects like aberrations in sleep, appetite, sexual desire, loss of pleasure in work, psychomotor slowing, and suicidal thoughts [1]. Risk factors associated with MDD include gender, traumatic life experiences, disruptive childhood events, particular personality traits, and substance misuse [3,4]. Seeing as many individuals harbor these risk factors, it is not surprising the wide scale to which depression affects so many individuals.

Symptom presentation, course of disease, and treatment response are all additional components of this disorder that can be characterized as heterogenous. Genome-wide association studies (GWAS) and candidate gene studies have found few robust and consistent genetic risk factors for MDD. The heterogeneity of MDD, and lack of evidence to guide personalized treatment strategies, makes MDD one of the most challenging chronic conditions to treat in addition to creating a quandary for biomedical researchers [5].

1.1.1 Epidemiology

Worldwide, depression affects more than 264 million people [6]. In the United States, the lifetime incidence of depression is 20% in women and 12% in men [7,8]. The larger impact on women is consistent when looking at prevalence as well. MDD affects women at a prevalence ratio of 5:2 [9]. The overall prevalence of MDD is only increasing as well [10]. Over a span of ten years from 2005 to 2015, prevalence has increased by 18.4% [11]. This increase is particularly concerning given that this disorder already causes significant disability, morbidity, and mortality worldwide. Specifically, MDD is the principal factor leading to suicide and disability from chronic illness in the world [6,12].

1.1.2 Genetics

Genetics have proven to be vital in working towards a greater understanding of multiple disorders. However, as mentioned previously, this avenue has not proven to be as fruitful in the case of MDD. The presence of many common genetic variants with small effect sizes being linked to MDD has been reported in numerous studies, namely the multitude of GWAS studies published before 2018 with zero likely variants uncovered [13]. Out of eight MDD GWAS studies conducted before 2018, only one locus was identified to have possible GWAS significance [14–21]. The validity of the locus 12q21.31, next to gene SLC6A15, being linked to MDD has been questioned due to the failure of subsequent replication studies [13,20]. Notably, the locus was not replicated in the 2018 GWAS that identified 44 independent and significant risk loci for MDD [22].

Despite the nascent and contradicting results of GWAS studies towards uncovering the molecular underpinnings of MDD, there have been some loci given special attention due to

purported associations. For example, rs2242446 is within a norepinephrine transporter gene and has been shown to be associated with 1.67-fold greater odds of symptom remission and a two week shorter time to remission in adults older than 60 years [23].

Another common polymorphic variant for MDD is 5-HTTLPR, which is in the proximal 5' regulatory region of 5-HTT and the promoter region of serotonin-transporter-linked gene SLC6A4. 5-HTTLPR modifies the promoter activity of the 5-HTT gene. Copy number variation of 5-HTT has been linked to depression symptoms, diagnosis, and suicide risk, where a short allele has lower transcription efficiency as compared to a long allele. Lower transcriptional activity results in less uptake of serotonin in the presynaptic neurons in the brain. Therefore, the short allele in 5-HTTLPR has been associated with more depressive symptoms [24,25]. Brain imaging studies have also shown functional differences in areas of the brain associated with 5-HTTLPR polymorphisms, which may be more broadly associated with complex traits and behavior, like depression [26,27].

In summary, a substantial majority of published literature surrounding this disorder have yielded negative findings [14–21]. Nonetheless, these negative findings are illuminating and contribute to the greater understanding of the underlying biology of MDD. While research on biological mechanisms and genetic underpinnings of MDD have shown variable results, treatment-centered studies have proved to be more informative.

1.1.3 Treatment

The main approaches to treating this common disorder can vary depending on disease severity and patient and provide preference; however, treatment strategies typically encompass lifestyle changes (exercise, nutrition, social support, sleep, stress reduction, etc.), psychotherapy

(cognitive behavioral, interpersonal, psychodynamic), electroconvulsive treatment, and/or medication. In patients with mild to moderate depression, psychotherapy options have been shown to have similar symptom remission rates to medication [28]. However, patients with moderate to severe depression often require treatment with antidepressants which may be lifelong. The current study will focus on medication as the primary MDD treatment modality of interest.

Pharmacologic treatment of MDD involves the prescription of antidepressants to ameliorate chemical imbalances in the brain through interactions with neurotransmitters. Neurotransmitters are held in vesicles within nerve cells or neurons. Serotonin, dopamine, and norepinephrine are all monoamine neurotransmitters that have been implicated in the pathophysiology of depression. These neurotransmitters are released by the presynaptic neuron into the synapse where they are able to interact with the postsynaptic neuron. Antidepressants can increase neurotransmission by increasing the release of neurotransmitters or by inhibiting the reuptake or degradation of neurotransmitters, thereby increasing the presence of neurotransmitters available in the synapse.

Antidepressants are often classified by their functional impact on neuronal synapses [3]. Selective serotonin reuptake inhibitors (SSRIs) are an antidepressant class that increase serotonin (5-HT) levels and activity in the brain by decreasing serotonin reuptake at synapses. Common SSRIs are citalopram (Celexa), escitalopram (Lexapro), fluoxetine (Prozac), paroxetine (Paxil, Pexeva), and sertraline (Zoloft). SSRIs are first-line treatments for depression, as supported by national and international clinical guidelines [29,30]. Therefore, SSRIs are the most commonly prescribed antidepressant. However, despite being supported as first-line treatment and prescribed frequently, SSRIs still exhibit relatively low success rates for patients. Only 35-45% of individuals experience symptom remission (a significant decrease in symptoms) with first-line treatment [31].

One study focusing on SSRI prescribing found symptom remission rates of 28% and response rates of 47% [32]. These relatively low frequencies may result in several weeks to months of troubleshooting antidepressant treatment options in a trial-and-error fashion before a successful drug trial is identified [32,33].

Other classes of antidepressants in the available drug arsenal for treating MDD include monoamine oxidase inhibitors (MAOIs), serotonin norepinephrine reuptake inhibitors (SNRIs), 5-HT2 receptor antagonists, dopamine reuptake inhibitors, and tricyclic antidepressants (TCAs). The first antidepressant class was MAOIs developed in 1952. MAOIs inhibit the breakdown of serotonin, dopamine, and norepinephrine into their metabolites, allowing more of these neurotransmitters to be available in the brain. Similarly to SSRIs, SNRIs also allow for more neurotransmitters to be available in the brain by inhibiting reuptake of serotonin as well as norepinephrine. SNRIs are typically recommended when patients do not respond to SSRIs [29,30]. Another antidepressant class is 5-HT2 receptor antagonists, which block serotonin reuptake through antagonizing 5-HT2 receptors. Dopamine reuptake inhibitors like bupropion inhibit dopamine and norepinephrine reuptake, and therefore represent a novel treatment option as they do not act directly on serotonin [34]. The last antidepressant class is an exception to antidepressant nomenclature where the name describes the functional neuronal synapse interaction. Tricyclic antidepressants (TCAs) are named for their characteristic chemical structure. TCAs like amitriptyline block reuptake of serotonin and norepinephrine. In comparison to other antidepressant classes, TCAs have lower 5-HT reuptake inhibiting effects, higher norepinephrine reuptake inhibition, and block several additional neurotransmitter receptors like those of alpha1 adrenergic and histamine resulting in a less favorable side effect profile [3,34].

Among the myriad of these aforementioned antidepressant treatment options, the prescribing possibilities are expanded further as antidepressants are prescribed in combination when a partial response in achieved. Response to antidepressant therapy is highly variable. This diversity in response to antidepressants has even been exhibited at the level of biological sex [35–39]. Despite this knowledge, no particular clinical variable or modality has been shown to be a consistent predictor for antidepressant therapeutic outcomes. In working towards informing antidepressant prescribing and advancing the current trial-and-error approach to treatment selection, there needs to be a method of determining which antidepressant a patient is more likely to respond to as well as systematic and standardized measurement-based symptom scoring to monitor treatment success, or lack thereof.

1.1.4 Methods of Measurement

There are many scoring metrics for MDD screening and assessing symptom severity. Currently on the fifth edition, the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) is a reference book that includes a classification system for MDD screening. The manual was developed by mental health professionals and organizations, including the American Psychiatric Association and the World Health Organization through comprehensive and systematic reviews of published literature, in addition to dataset and field trial analyses. DSM-5 builds on its previous versions and has a strong foundation in empiricism with consensus support from the mental health field consensus. The DSM-5 scoring criteria was used in the development of the International Statistical Classification of Diseases and Related Health Problems (ICD-10), and thus is directly related and historically tied to the foundation of MDD diagnosis [40].

After the publication of the first DSM in 1952 for detecting and diagnosing depression, the Hamilton Depression Rating Scale, HAMD, was created to assess patients already diagnosed with MDD. This rating scale is one of the most commonly implemented instruments for quantifying symptoms and gauging symptom severity during a patient encounter [41,42]. The rating scale includes measurements for 17 variables (low mood; guilty feelings; suicidal thoughts; insomnia; decreased engagement in work and interests; retardation; increased agitation; psychic and somatic anxiety; general, gastrointestinal, and genital somatic symptoms; hypochondriasis; deeper insight; and loss of weight) and has been demonstrated to be reliable and consistent [41,42].

Another depression scoring metric is the Patient Health Questionnaire (PHQ), which is a self-administered survey that is commonly used to assess depressive symptoms in clinical practice. The PHQ-9 includes nine criteria from DSM-5 and has been shown to be a reliable and valid depression scale that can be administered in one to five minutes, whereas other scoring metrics can take up to 20 to 30 minutes to complete [43]. Shorter versions of the PHQ are also available including the two-item PHQ-2 and four-item PHQ-4.

Another depression scoring metric is the Inventory of Depressive Symptomatology (IDS), along with the Quick Inventory of Depressive Symptomatology (QIDS). These metrics are 30- and 16-items, respectively, used to quantify depression symptom severity. Both scoring metrics have been shown to be sensitive and acceptable measurements of MDD symptom severity [44]. These two scoring metrics use a 4-point Likert scale focused on behaviors and moods from the previous week. Additionally, there are variants of the QIDS metric that allow for adaptability in terms of the survey administer. The QIDS-C16 is a clinician-rated version, as specified by the “C”. The numeric value stands for how many items are on the survey. The patient-rated version of QIDS is QIDS-SR, as denoted by the “SR” for self-rated. QIDS-SR has been demonstrated to be as

sensitive to symptom changes as the IDS-SR_{30} and $HAMD_{24}$ metrics, and thus is a sufficient symptom scoring tool [45].

Like the number of antidepressant options, there are a multitude of depression scoring metrics with different strengths and limitations. Selection of the optimal scoring metric for a given scenario depends on its flexibility to administer, length of time needed, and survey structure. Scoring metrics are one piece of a broader puzzle working to better understand MDD presentation and treatment response.

1.1.5 Randomized Clinical Trials

Randomized clinical trials (RCTs) using depression scoring metrics to quantify clinical endpoints of treatment response or symptom remission have demonstrated efficacy of individual antidepressants [46–48]. However, the findings of these RCTs have been cited to not be generalizable to clinical practice due to RCTs exclusion of psychiatric comorbidities [49]. In addition, RCTs typically only compare only a few antidepressants per study for effectiveness [48]. In response to the lack of generalizability to a real-world clinical population and in an effort to compare effectiveness of many antidepressants in a study that follows a clinical workflow prescribing, the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial was designed [49].

The STAR*D trial was a large multi-center trial funded by the National Institutes of Health (NIH) that sought to address antidepressant treatment effectiveness in MDD patients [50]. This study has the largest sample population and the longest study duration to assess depression treatment outcomes. The goal was to assess depression treatment effectiveness and tolerability for MDD patients [51]. Given that two thirds of patients do not experience symptom remission after

first-line treatment, researchers wanted to identify alternative treatment strategies to help this majority of MDD patients. There were 4,041 patients enrolled, and almost half of study participants (N=1,948) were sequenced.

The study was designed to reflect clinical practice, with patients started on the baseline MDD treatment of citalopram (Celexa) and given the choice between a list of antidepressants that they would be open to trying to treat their depression. If patients continued to experience MDD symptoms after 12 to 14 weeks, they proceeded to the next level of the trial where patients were given the choice to add to their current treatment, switch to an antidepressant at random from the list of antidepressants, and/or add cognitive psychotherapy as part of their treatment. There were four levels total in the study. Patients were classified to have highly treatment-resistant depression at level four.

The STAR*D trial lasted for seven years and enrolled adults from age 18 to 75 years old at 41 clinical locations across the U.S. Overall, STAR*D found that 33% of patients achieved symptom remission, and 10-15% more experienced symptom response (symptoms decreased by at least half), but were not symptom-free [50]. On average, it took patients six weeks to achieve symptom response, and seven weeks to achieve symptom remission. About half of the patients achieved symptom remission after level two. Over all four levels, almost 70% of study participants achieved symptom remission, however drop-out was meaningful across the four levels. Overall, STAR*D was a well-designed study that led to 124 publications, as reported by clinicaltrials.gov (Accessed 6/20/20). Most notably of these publications, studies identified multiple genes of interest associated with treatment response: HTR2A, GRIK4, SLC6A4, FKBP5, and TREK1 [25,51–53]. These uncovered associations between gene variants and response to antidepressant treatment were valuable pharmacogenomics findings.

1.2 Pharmacogenomics

Pharmacogenomics (PGx) is the study of how genomic variation impacts an individual's response to medication. It is the application of genomic data to inform drug and dose selection. Therefore, PGx is a combination of pharmacology and genomics. It is important to know the implications of gene-drug pairs because some combinations can cause patients to experience varying effectiveness or, can even be life-threatening.

Pharmacogenetics, the study of individual gene interactions with drugs, was first introduced in 1959 by Friedrich Vogel. With new developments in sequencing technology and the greater capacity to understand many genes at once, genomics came to the forefront. With it, came the study of multiple gene-drug interactions or, PGx. PGx as a field aims to inform prescribing to maintain drug efficacy and minimize adverse drug reactions.

The largest research network in this space is the Pharmacogenomics Research Network (PGRN). The network is comprised of three research centers and two public PGx resources (Pharmacogenetic Knowledge Base and PGRN Hub) and states that its mission is to fuel precision medicine by supporting discovery and translation of genomics that informs therapy and adverse drug reactions [54,55]. The network works to empower this mission through promoting PGx research and advising the clinical sphere on the importance of PGx.

PharmGKB is the Pharmacogenetic Knowledge Base which hosts genomic, phenotypic, and clinical data from pharmacogenomic studies that can be browsed and queried [56]. Created out of PharmGKB is the Clinical Pharmacogenetics Implementation Consortium (CPIC), an international academic consortium of individual researchers and staff members established in 2009 under a commitment to guide the use of pharmacogenetic test results to inform patient care [57]. CPIC publishes clinical practice guidelines that are peer-reviewed and evidence-based in effort to

improve implementation of PGx tests in clinical practice. CPIC hosts 84 practice guidelines (Accessed 6/15/2020) [58]. Two of these guidelines are dedicated to recommendations for genes that play a role in antidepressant metabolism, specifically SSRIs and TCAs [59,60]. The Food and Drug Administration also lists PGx biomarkers on approved drug labels, and recognizes PGx associations along with CPIC [61,62].

1.2.1 Research Studies in Pharmacogenomics

There are 127 unique genes and 240 unique drugs for which CPIC lists gene-drug associations. In total, there are currently 377 gene-drug pairs noted by CPIC, however there is variable evidence assigned to each pair. A few of the most prominently known PGx genes are *CYP2D6, G6PD, CYP2C9, CYP2C19, ABCB1*, and *HLA-B* (Table 1). The number of drugs reported to have any associations with these genes are noted in the second column of Table 1, however CPIC specifies that only gene-drug pairs that have published recommendation guidelines have undergone an appropriate evidence review to provide definitive CPIC recommendations for prescribing. CPIC rates the degree of evidence in support of gene-drug pairs from levels A to D. Likewise, PharmGKB rates evidence as 1A, 1B, 2A, 2B, or 3. While many gene-drug associations have been reported in the literature, a smaller number of gene-drug combinations have shown high levels of evidence.

Table 1 Prominent PGx genes, the number of genes they are reported to be associated with, and the level of evidence found in support of the gene-drug pair.

Most prominent PGx genes	Number of drugs listed by CPIC as having a genetic variation-drug response association	Number of drugs reported to have variable effectiveness based on genotype, with Level A CPIC evidence	Number of drugs reported to have variable effectiveness based on genotype, with Level 1A PharmGKB evidence
CYP2D6	60	12	23
G6PD	36	2	2
CYP2C9	22	10	10
CYP2C19	21	5	13
ABCB1	12	0	0
HLA-B	11	5	11

1.2.2 Implementation in Clinical Care

Gene-drug pairs and PGx variants are vitally important to clinical knowledge because over 50% of all primary care patients are at some point exposed to PGx relevant medications [63]. More so, 18% of the total prescriptions written in the U.S. per year are affected by actionable PGx variants [64–67]. In addition, common genetic variation attributes to 42% of antidepressant response variation [68]. Therefore, response to antidepressant treatment is complex and related to many common genetic variants with small effect sizes [69]. There has been significant work dedicated to making PGx knowledge more accessible to clinicians through clinical decision

support (CDS) tools and educational programs. Many institutions, including the University of Pittsburgh Medical Center, are working to roll out PGx CDS alerts to prescribers when appropriate [70,71]. In addition, significant work has been dedicated to rigorously funded and researched Continuing Medical Education programs that increase clinical knowledge surrounding PGx [72].

Not only do clinicians need to be aware of PGx variants and associated drugs because the variants and drugs are so prevalent, but also because informing care based on PGx has been cited for its potential to decrease healthcare costs through reducing adverse drug reactions, failed clinical trials, time to drug approval, length of medication durations, number of medications prescribed, and the effects of disease on patients [73].

Despite these advantages to incorporating PGx knowledge into clinical decision making, there are major undeniable barriers to clinical implementation. In the information technology sphere, there are challenges associated with gathering and storing data, ensuring data security, developing an efficient CDS infrastructure, optimizing user interface principles, minimizing alert fatigue, and working with currently unstandardized terminologies and language surrounding PGx [73,74]. These challenges have not been prohibitive to researchers implementing PGx into the clinical space, as exemplified by the eMERGE-PGx project rolling out PGx data in electronic health records (EHRs) [75]. However, clinical implementation of PGx in practice has had slow buy-in from clinicians due to a lack of evidence supporting routine PGx testing, a dearth of randomized controlled trials (RCTs) supporting gene-drug associations, a myriad of genotyping tools with variable, uninteroperable outputs, as well as a lack of an established presence in the current clinical workflow. Also, as alluded to previously, there is a general lack of PGx knowledge or appreciation among clinicians regarding utility of PGx in clinical decision making. To further complicate PGx implementation, insurance reimbursement for PGx testing is variable, leaving

many patients paying high out-of-pocket costs [73]. Despite these hurdles, PGx research has managed to demonstrate clinical utility in specific areas of practice.

One such important PGx implementation in clinical settings is *CYP2C19* genotyping. *CYP2C19* variants predict clopidogrel response and therefore inform antiplatelet treatment after a percutaneous intervention (PCI) [76]. The CYP2C19 gene is crucial in transforming clopidogrel to becoming pharmacologically active. A nonfunctional variant can be lethal due to fewer active clopidogrel, which can lead to increased risk for major adverse cardiovascular complications following a PCI. Sixty five percent of Asians and 30% of whites have nonfunctional CYP2C19. University of Florida initiated *CYP2C19*-guided prescribing in 2012, and has led the way in informing *CYC2C19*-guided therapy [77,78]. Other clinical implementations of important PGx variants in care are *CYP2D6* with codeine, tramadol, opioids, SSRIs, aripiprazole, and atomoxetine; *TPMT* with thiopurines; *CYP2D6* and *CYP2C19* with SRRIs and TCAs; *CYP2C19* with PPIs, voriconazole, and citalopram; *SLCO1B1* with simvastatin, among many others [76].

1.3 PGx and MDD

PGx-guided therapy has been shown to be important for symptom remission and response, and therefore, informative for improving patient outcomes. Hall-Flavin et al. (2012) found that the implementation of a PGx algorithm to guide depression treatment resulted in a statistically significant reduction in depression symptoms, quantified by QIDS-C16 and HAM-D17 scores, compared to patients treated without the PGx algorithm [79]. Bradley et al. (2018) also found that drug response and symptom remission rates were higher at 12 weeks among patients treated under a PGx-guided group compared to those that were not [80]. Multiple systematic reviews have also

replicated these findings of improved symptom remission in PGx-guided patients. Bousman et al. (2019) looked at 1,737 subjects from five RCTs, and found that patients in the PGx-guided group were 1.71 times more likely to achieve symptom remission [81]. The researchers went on to recommend that PGx-guided therapy be implemented in the clinic. Additionally, Rosenblat et al. (2017) also found from a systematic review of clinical trials that PGx testing could improve symptom remission [82]. In addition to symptom remission findings, studies have also shown that PGx data is vital in identifying poor metabolizers of antidepressants due to their adverse effects [83,84].

Stemming from these results highlighting the importance of PGx-informed therapy for MDD, there are two CPIC guidelines published with moderate to high quality of evidence surrounding SSRI recommendations for *CYP2D6* and *CYP2C19* genotypes, and high quality evidence in most cases for TCA recommendations based on *CYP2D6* and *CYP2C19* genotypic variations [59,60]. It is also meaningful to consider that the results of studies demonstrating symptom remission and response have been criticized for their study design, sample size, and funding sources. Namely, funding sources for prominent antidepressant studies have been sourced from pharmaceutical companies that have reported conflict of interest [80,85,86]. There still is a limited evidence base for using PGx to inform antidepressant prescribing, and therefore a need to support the clinical utility of PGx-guided therapy reliably and robustly for MDD outcomes.

In effort to work past this limited evidence base, and in consideration of the magnitude and expense of conducting RCTs, there needs to be other approaches to gathering evidence on PGx testing outcomes besides an RCT for every PGx relevant gene-drug pair [76]. One avenue that has been pursued is whether real-world evidence from EHRs can be used to evaluate and replicate RCTs using their interventions, inclusion/exclusion criteria, and primary endpoints [87]. This real-

world data affords a more cost-effective solution and would facilitate more discovery from already existing EHR data. Although, there are notable limitations here as well in that the observational and incomplete nature of EHR data is problematic in terms of establishing causal effects.

1.4 PGx and EHRs

The EHR is a valuable real-world data resource, and amid the implementation of PGx data into clinical care through the EHR, we can additionally make use of both EHR and PGx data to inform models. The Electronic Medical Records and Genomics (eMERGE) network is particularly valuable to this endeavor, in that it is an NIH-organized and funded institution that supports and encourages the combination of biorepositories with EHR systems in order to allow for research surrounding genomic discovery and genomic medicine.

The eMERGE network and PGRN partnered together in a leadership effort to burgeon the field of genomic discovery and medicine using PGx and EHR data. The eMERGE-PGRN project has demonstrated the implementation of genetic sequence data (84 PGx candidate genes from 9,000 participants) into clinical practice to inform prescribing through CDS [75,88]. Despite this leadership effort, it is no surprise that based on the nascence of evidence for the clinical utility of PGx, the lack of PGx knowledge within the prescribing community, and PGx testing not being established yet in routine clinical care, that having PGx data within EHRs is not standard. However, progress is being made in many health systems, and we are on the cutting edge of using PGx data to inform clinical care.

1.5 PGx and Machine Learning

Machine learning is a favorable approach to uncover relationships in data without explicitly specifying the model equation. ML models are useful to sort through large datasets where there are unknown relationships between feature inputs, and so the algorithm can uncover variables that are informative to predicting outcomes.

In the case of using high-dimensional PGx data along with high-dimensional EHR data, machine learning (ML) is an ostensibly favorable approach due to its' ability to uncover underlying patterns and relationships in large datasets that are unknown or difficult for humans to discern [89]. There have been many publications outlining the promise and value of this work uncovering relevant features to improve antidepressant treatment response prediction in MDD patients [90–92]. However, there is still significant contributions necessary to turn the promise of this work into a reality, and to ultimately translate predictive models into clinical care.

One of the earliest and most notable publications that used ML methods focused on predicting antidepressant treatment outcome was by Chekroud et al. (2016) [93]. Researchers used STAR*D trial level 1 data consisting of around 4,000 patients to predict symptom remission to citalopram after 12 weeks. From 164 clinical variables, they found 25 relevant features for prediction that surrounded somatic complaints, insomnia, and traumatic life experiences. Their model was able to predict symptom remission at an accuracy of 65%.

Another application of ML methods to early depression was from Kautzky et al. (2015), which investigated 225 patients, 98.7% of which were Caucasian, from the European Group for the Study of Resistant Depression [94]. Kautzky et al. ran random forests to perform feature selection and performed k-means clustering to determine features that were associated with treatment outcomes. They looked at 12 SNPs in five genes, eight clinical variables, and used HAM-

D to measure symptom progression. Researchers ultimately identified 3 SNPs and one clinical variable that was significantly associated with treatment response. However, their reported sensitivity for identifying patients with treatment response within 74 patients was low at 25%. Their use of random forests for variable selection, limited features in their dataset, small, homogenous study population, and extremely low sensitivity left room for improvement.

Athreya et al. (2018) demonstrated much better performance in predicting antidepressant treatment outcomes with an accuracy of 80% [95]. To achieve this, Athreya et al. used physician assessments along with metabolomic, genetic, and sociodemographic data from 603 patients to predict antidepressant response. Patients were treated with citalopram for 8 weeks, over which researchers assessed symptom severity at zero, four, and eight weeks after initial treatment through QIDS-C. This trial achieved sufficient prediction accuracy and demonstrated a strong study design, however the model included physician assessments, which limited generalizability.

In 2019, Athreya et al. addressed the issue of including physician assessments, and expanded upon previous work to focus on using PGx data specifically for treatment outcome prediction [96]. The study population was 1,030 white outpatients that were genotyped for four genes and two metabolite concentrations, serotonin and kynurenine. Researchers subset their models based on sex due to demonstrated differential responses to antidepressants. The supervised ML model trained on SNPs and baseline depression scores was able to predict symptom remission and treatment response at 8 weeks with an area under the receiver operating characteristic curve (AUC) of 0.7 and an accuracy of 0.69. Performance notably decreased when researchers did not include physician assessments, however performance was still impressive.

Moving forward with more high-dimensional PGx data and clinical biomarkers, Lin et al. (2018) implemented a deep learning method using demographic information, SNP data, baseline

HAMD scores, depressive episodes, and suicide attempt status from 455 patients to classify antidepressant responders [97]. They used a multilayer feedforward neural network with two hidden layers to achieve an AUC of 0.83 in predicting antidepressant response. For predicting symptom remission, they had an AUC of 0.81 with three hidden layers. This study demonstrated the feasibility of implementing deep learning frameworks to predict both treatment response and remission.

With a number of successful prediction studies, Perlman et al. (2019) sought to characterize the feature space of ML studies predicting antidepressant treatment outcomes [98]. They noted around 200 features that were important inputs for depression models. Also taking a comprehensive approach to understanding the scope of depression ML studies, Lee et al. (2018) conducted a meta-analysis and systematic review and found that classification algorithms performed better with multiple high-dimensional data types, and that models classifying treatment outcomes had an average accuracy of 0.82 overall [99]. There are many research groups looking to improve treatment outcome prediction for antidepressants, but this is in no way a solved issue. Prediction stands to be improved, methods stand to be more robust, reproducible, and generalizable, and study populations stand to be larger and more diverse.

We seek to address these limitations of previous studies through the use of the Pitt + Me Discovery cohort. This cohort will ultimately enroll 150,000 patients with PGx testing for 4,627 markers within 1,191 genes, and their EHR data, including clinical notes. We plan to make use of established electronic phenotypes in order to proxy depression scoring metrics and enhance MDD phenotyping past ICD-10 classification. Structured EHR data will also greatly improve our methods over previous ML depression studies as well. Few studies have optimized on the juncture between using EHR and PGx data in order to improve antidepressant treatment response

prediction. With the mission to improve on previous models' limitations, we remember George Box's famous quotation:

"All models are wrong, but some are useful" – George Box

With this quotation in mind, we proceed with caution, but also inspiration into the space we have carved out to improve antidepressant treatment response prediction for MDD patients.

2.0 Research Design

2.1 Overall Design

This study is a secondary analysis of existing electronic health record (EHR) cohort data on patients with major depressive disorder. EHR data was pulled through the Health Record Research Request (R3) team from University of Pittsburgh Medical Center data. Inclusion and exclusion criteria were established for the original cohort pull based on specified ICD-9 and -10 codes from eMERGE's PheKB database for depression. In addition, patients had to be 18 years of age or older at the time of the data pull. Earliest records pulled were from 2004, being that data quality significantly decreased before 2004. Exclusion criteria surrounded ICD-9 and -10 codes for disorders that may also be prescribed antidepressants though may contribute confounding, for example bipolar disorder, anxiety disorder, post-traumatic stress disorder, autism, etc. For this study, we only looked at outpatient data for the intended capture of low severity major depressive disorder cases, as opposed to inpatient psychiatric data that may be complicated by additional comorbidities.

In addition, the reproducible data analysis and machine learning pipeline was designed for the eventual incorporation of PGx data from Pitt + Me Discovery. Pitt + Me Discovery is a biorepository within the Pitt + Me clinical trial enrollment system that seeks to collect PGx data from 150,000 patients of 18 years of age and older. Pitt + Me Discovery was initialized under the objective of learning from patients' genetic data to further understanding of health, disease outcomes, and patient response to medication and treatment. Another objective of Pitt + Me Discovery is to communicate information gleaned from using PGx data to inform prescribing. PGx

on MDD patients from within the Pitt + Me Discovery cohort may be useful to improve antidepressant prescribing. Data from R3 included patient demographics, diagnoses, encounters, medication orders and fills, psychiatric questionnaires, and vitals for all UPMC patients that fell within the inclusion criteria for MDD.

2.2 Specific Aims

In effort to predict antidepressant treatment response and symptom remission for MDD patients, we developed a comprehensive and systematic data analysis pipeline accomplished through the following aims:

Aim 1. Extract and process electronic health record (EHR) and implement established approaches for electronic phenotyping of depression status.

Aim 1a. Extract, process, and perform data quality checks on UPMC EHR data.

Rationale: Missing, incomplete, and messy data are inherent to electronic health record data. Therefore, data processing and quality checks are vital to minimizing bias and improving accuracy measures in downstream analyses.

Hypothesis: Data processing and quality checks ensure a complete analysis and improve specificity in examining a cohort of MDD patients.

Approach: Summary statistics were reported, data was assessed for level of missingness, and imputation was determined to be unnecessary due to a low degree of missingness.

Aim 1b. Implement established electronic phenotyping for identifying patients with depression.

Rationale: eMERGE has an established and validated electronic phenotype for depression, which allows for increased specificity in identifying a cohort containing patients with depression.

Hypothesis: The implementation of the eMERGE phenotype generates greater specificity for depression patients, which improves downstream model performance, as measured through F1 scores and accuracy.

Approach: In order to address this aim, the inclusion and exclusion criteria of ICD-9/10 codes from eMERGE were applied. Additionally, the "2/30/180 rule" of evidence of depression present on two different calendar days, at least 30 days apart, and no greater than 180 days apart, was implemented.

Aim 2. Characterize antidepressant prescribing sequences and formulate (and evaluate the accuracy of) a baseline clinical model.

Aim 2a. Describe antidepressant prescribing sequences and model associated transition probabilities with Markov Models.

Rationale: Markov Models were applied to the dataset in order to illustrate patients' progression through antidepressant treatment sequences.

Hypothesis: Markov Models convey the large proportion of patients that are prescribed SSRIs initially, and then disperse to a number of antidepressant treatment sequences.

Approach: Markov Models were created for medication fills using the "markovchain" package in R [100]. Graphics were constructed to illustrate progressions through medication states and transition probabilities.

Aim 2b. Use antidepressant prescribing sequences and patient characteristics from the EHR to formulate and evaluate the accuracy of baseline models (logistic regression models, classification trees) for predicting treatment response and symptom remission.

Rationale: Statistical and machine learning models were constructed to ascertain features associated with depression patients' symptom remission and treatment response to antidepressants, and predict remission and response to antidepressants.

Hypothesis: Statistical and machine learning models are able to model and predict treatment response and symptom remission to antidepressants with greater accuracy than what is currently observed clinically (about 50%).

Approach: Traditional statistical models (e.g., logistic regression models) and machine learning models (e.g., random forest) were fit to examine treatment response and symptom remission.

Aim 3. Design models for the eventual incorporation of PGx data to the baseline clinical models to eventually evaluate the value-added of PGx data to statistical models, and develop, implement, and evaluate accuracy of a reproducible machine learning pipeline to predict treatment response and symptom remission.

Aim 3a. Incorporate PGx data in the theoretical clinical model architecture as a part of the reproducible pipeline to be eventually run when there is sufficient PGx data, and subsequently evaluate predictive accuracy changes.

Rationale: PGx data is an eventual added feature to the baseline clinical model. The statistical and machine learning models are run and analyzed again to compare resulting model performance, in terms of accuracy and F1 score.

Hypothesis: PGx data provide additional features in the training set that eventually enhance symptom remission and treatment response prediction.

Approach: PGx data can be included in the statistical and machine learning model architectures to be run in the future to allow for the downstream evaluation of accuracy and F1 score compared to the baseline clinical models naïve to PGx data.

Aim 3b. Develop, implement, and evaluate accuracy of a reproducible analysis pipeline (logistic regression, random forests, and ensemble machine learning methods) to predict treatment response and symptom remission.

Rationale: A reproducible analysis pipeline was constructed to allow for analyses to be reproduced and expanded as additional patients are enrolled in the Pitt + Me Discovery cohort. More so, a reproducible pipeline allows for a favorable opportunity to generalize findings to similar clinical datasets.

Hypothesis: A reproducible pipeline allows for reuse, modifications, tuning, further testing, and transport to adjacent datasets and hospital settings.

Approach: Methods were automated so that inputs and outputs of steps follow seamlessly in an automated and generalizable fashion. Future directions could involve mobilizing this pipeline into a Docker container for ultimate reproducibility, transportability, and generalizability.

We hypothesize that our novel analysis pipeline and patients' clinical care features from the EHR demonstrates enhanced prediction of treatment response and symptom remission to antidepressant treatments.

3.0 Data Acquisition and Manipulation

3.1 Introduction

3.1.1 Data source

University of Pittsburgh Medical Center (UPMC) is a $21 billion health care provider and insurer in Pittsburgh, Pennsylvania. There are 40 academic, community, and specialty hospitals, and 700 doctors' offices and outpatient sites in the UPMC system that are located across Pennsylvania, New York, and Maryland. UPMC medical sites host specializations in transplantation, cancer, psychiatry, neurosurgery, geriatrics, rehabilitation, and women's health.

In this study, we harnessed outpatient data in the UPMC EHR in order to capture depression and its treatment in ambulatory patients as opposed to inpatient data which may be populated by more complicated and less generalizable cases. Outpatient data included EHR data from numerous outpatient sites in the UPMC hospital system.

3.1.2 UPMC Electronic Health Record Data

UPMC uses two electronic health record (EHR) providers: Cerner for inpatient service and Epic for outpatient. In this study, only Epic data was accessed. To allow for use of this EHR data for research purposes, the University of Pittsburgh constructed Neptune, a data warehousing resource (Figure 1). Neptune pulls EHR data from Cerner and Epic periodically to provide researchers with de-identified data. The data is delivered through a consult and honest broker

service called the Health Record Research Request (R3), within the Department of Biomedical Informatics (DBMI) at the University of Pittsburgh.

The data for this study was acquired through these mechanisms, after being approved through the University of Pittsburgh's Institutional Review Board under an Exempt criteria for secondary research on data or specimens (STUDY20020047).

3.1.3 Future Data Collection from the Pitt + Me Discovery Data

The Pitt + Me Discovery is a biorepository that seeks to enroll 250,000 patients that are 18 years of age and older. The Pitt + Me Discovery cohort was created to learn from patients' genetic data in order to gain a better understanding of health, disease outcomes, and patient response to medications and treatment. Insights from this genetic data are expected to improve prediction, diagnosis, treatment, and prevention of diseases.

The Pharmacogenomics Center of Excellence, a joint venture of PittPharmacy and Thermo Fisher Scientific, seeks to use the pharmacogenomics panel conducted on patients within Pitt + Me Discovery to demonstrate the value of pharmacogenomics implementation in clinical care. The pharmacogenomic panel is tested on patients' blood or saliva samples and consists of 4,627 markers within 1,191 genes.

A future application of this work will be to demonstrate the information gain associated with the addition of pharmacogenomics data when making prescribing decisions. Data from the Pitt + Me Discovery cohort will ultimately be included in constructed models to test the value added of pharmacogenomics data. However, enrollment was slowed during the COVID-19 global pandemic and therefore, there was not sufficient PGx data to be included in this study.

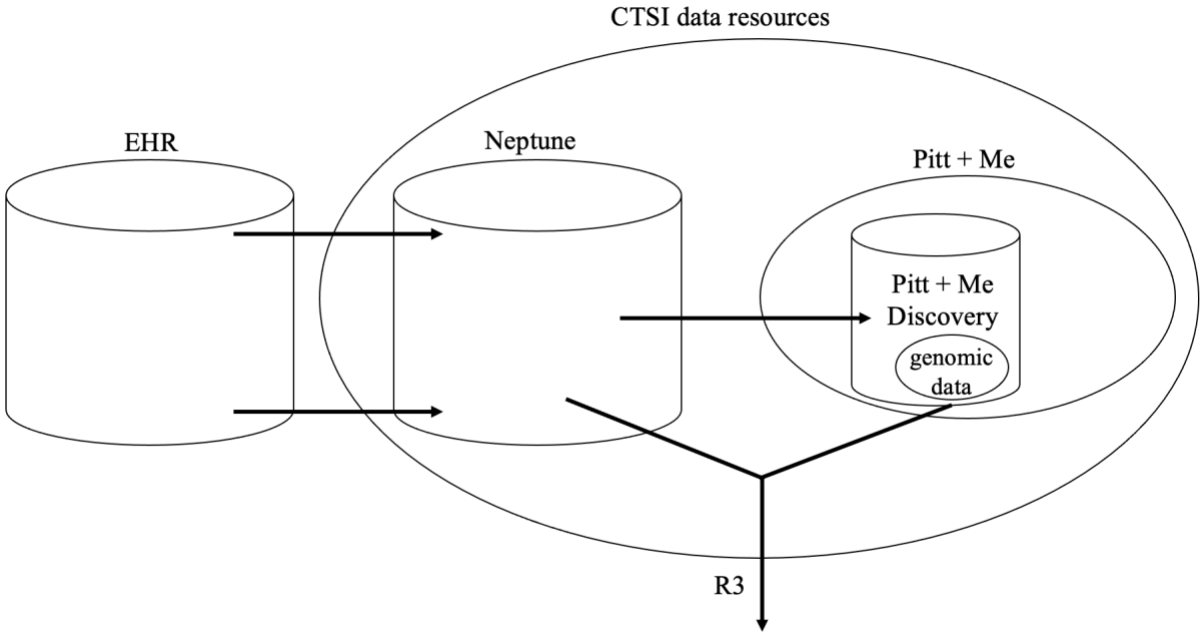


Figure 1 Data warehouse infrastructure.

3.1.4 Study Dataset

The study dataset consisted of patients within the UPMC system that satisfied the inclusion criteria. The inclusion criteria were based on the PheKB depression definition, which involves ICD-9/10 codes for major depressive disorder and/or the prescription of an antidepressant. Patients were also included if they were 18 years of age or older at the time of the data pull. Exclusion criteria consists of a list of ICD-9/-10 codes for concomitant disorders that may also be prescribed antidepressants though may contribute confounding, for example bipolar disorder, anxiety disorder, post-traumatic stress disorder, autism, etc. (Appendix Table 1).

For UPMC patients that fell within the inclusion criteria, EHR fields consisted of patient demographics (gender, date of birth, race, patient status, death date), encounters with clinical care

facilities (admission date, discharge date, encounter type, facility/department, admit source, admitting diagnosis, primary diagnosis, Diagnosis-related group code), medication orders (medication, RxNorm code, order date, quantity, refills, start date, end date, instructions, pharmacy class, simple generic code), medication dispense information (medication, drug name, National Drug Code, dispense date, amount, quantity, frequency), diagnosis (diagnosis code, diagnosis name, diagnosis from date, diagnosis to date, primary diagnosis indicator), vitals (weight, height, contact date), and clinical notes (Table 4). Earliest records in the dataset are from 2004.

3.2 Methods

3.2.1 Data Processing and Missingness Assessment

Data was analyzed for each features' degree of missingness. Missingness is inherent to electronic health record data, and the proper handling of missingness is vital to minimizing bias in clinical dataset analyses. However, decisions surrounding how to handle missing data vary according to context. A common approach to handling missing data is to omit features and observations with missingness, which is called "complete case analysis". Alternatively, data handlers may decide to conduct imputation in order to insert values that were missing.

Mean, median, or mode-based imputations tend to result in the same estimate as the complete case analysis, but can also increase bias and underestimate variance [101]. In single imputation, a single rule replaces missing values for a particular feature. In multiple imputation, many possible datasets for missing values are created based on the selected imputation rule, an estimation step is conducted, and results are pooled to create one complete dataset resulting from

the multiple imputation [102–104]. Multiple imputation using chained equations (MICE) is a popular imputation method that regresses on all features to estimate imputed features' posterior distributions [105].

Given many features with imbalanced data, Synthetic Minority Oversampling Technique (SMOTE) is another imputation technique [106]. Other possible imputation methods are implementing clustering techniques to determine whether patients cluster according to certain features allowing the missing value to be deduced. Another method to deal with missingness might be to treat values as a separate category. Single (mean, median, or mode) imputation can also be performed within a feature across patients, or multiple imputation across many features. Features can have values that are missing at random (MAR), missing not at random (MNAR), or missing completely at random (MCAR). MNAR and MAR mechanisms are impossible to uncover based on observed data, though MAR is unlikely in the clinical setting [107,108]. Missing data can be classified to be MCAR if there missing values are independent of both observed and missing values [107,108].

3.2.2 Defining Analysis Cohort for Depression

Once a complete case analysis dataset was established, the dataset was analyzed to ensure the examination of patients with early depression. The electronic medical records and genomics (eMERGE) network rule-based electronic phenotype for MDD was additionally implemented.

eMERGE established a "2/30/180" rule, which states that evidence of depression must be present on two different calendar days, at least 30 days apart, and no greater than 180 days apart. This rule is important in order to rule out administrative artifacts and errors, and also allows one

to assume that medications that are ordered or dispensed within 30 days of an MDD electronic phenotype ICD-9/10 code are likely for treating MDD.

3.2.3 Description of EHR fields

For each UPMC patient that fulfills the establish inclusion and exclusion criteria for depression, EHR fields were extracted into a database (Table 2). The demographic table held patients' gender, date of birth, race, patient status, and death date (if applicable). The encounter table held information surrounding each of the patients' outpatient encounters, namely, the admission date, discharge date, encounter type, admitting diagnosis, primary diagnosis, diagnosis related group (DRG), and the PHQ score. The medication order table included the medication, RxNorm code, order date, number of refills, start and end date of the prescription, instructions, pharmacy class, and simple generic code. The medication dispense table will hold similar information: medication, drug name, national drug code (NDC), dispense date, amount, quantity, and frequency. There was also a table for diagnoses containing the diagnosis code, diagnosis name, diagnosis start and end date, and primary diagnosis indicator. Another table held vitals information, with height, weight, and contact date.

Table 2 Clinical Variables in the EHR database

Type	Fields
Demographics	Study ID
	Birth Date
	Death Date
	Gender
	Race
	Ethnicity
	Adult Consent Status
Encounter	Admission date
	Discharge date
	Encounter type
	Admitting diagnosis
	Primary diagnosis
	Diagnosis related group (DRG)
	Patient health questionnaire (PHQ) score
Medication order	Medication
	RxNorm
	Order date
	Refills
	Start date
	End date
	Instructions
	Pharmacy class
	Simple generic code
Medication dispense	Medication
	Drug name
	National drug code (NDC)
	Dispense date
	Amount
	Quantity
	Frequency
Diagnosis	Diagnosis code
	Diagnosis name
	Diagnosis from date
	Diagnosis to date
	Primary diagnosis indicator

Table 2 (Continued)

Type	Fields
Depression Questionnaire	Study ID
	Contact Date
	Form ID
	Form Name
	Questionnaire ID
	Questionnaire Name
	Question
	Answer ID
	Answer
Vitals	Weight
	Height
	Contact date

3.2.4 PHQ score analysis

PHQ scores were captured from the Questionnaire table in the EHR dataset. Multiple versions of the PHQ exist and are utilized at UPMC. The PHQ-2 is a two-item measure that is often utilized as the initial clinical screen for depression and is coded when a patient does not screen positive for depression. The PHQ-4 is a four-item measure that is used to screen for both depression and anxiety. The PHQ 2-8 is coded when a patient screens positive for depression, and therefore is subsequently asked questions three through eight. In some cases, a patient may be given the full nine-item PHQ questionnaire at their appointment; however, the PHQ-9 only administered in person due to the fact that the final question asks about suicidality. The PHQ-8 does not ask about suicidality and is therefore better suited to be administered remotely in the case that suicidality is missed and not responded to in real time.

PHQ-8 scores for patients over time were captured from the EHR and summary statistics were calculated. For patients with greater than one PHQ-8 score within the EHR, slopes of PHQ-

8 scores were calculated between patients' first and last PHQ-8 score. PHQ-8 scores for patients over time and PHQ-8 score slopes were plotted in boxplots to compare distributions of PHQ-8 scores and slopes according to subsets of patients' race and gender. One-way analysis of variance (ANOVA) tests were computed for group means of PHQ-8 scores and PHQ-8 score slopes between race, gender, and both gender and race together for patients. A Tukey's Honest Significant Difference (HSD) test was used to compare group means of PHQ-8 scores between races, and gender and race together. PHQ-8 score distributions for gender and race patient subsets were tested for normality using the Shapiro-Wilks test.

3.3 Results

3.3.1 Data Description, Summary Statistics, and Missingness Analysis

3.3.1.1 Demographics

Table 3 shows demographic data of each cohort. The gender identity distribution of this dataset parallels that observed in the general population, with most major depressive disorder patients identifying as female. This dataset is made up of 74.2% females, and 25.8% males (Table 3). Less than 3% are under 20, and less than 6% were over 84. Almost half (49.8%) of individuals are less than 55 years old. Most (93.2%) patients are presumed alive today. Additionally, most (92.3%) patients in the dataset are Caucasian. Only 5.3% of individuals are Black and 0.86% are Asian. While 0.9% of individuals in the dataset identify as Hispanic or Latino, most (94.2%) identify as "Not Hispanic or Latino".

Table 3 Demographics of cohort breakdowns, including demographics of patients excluded based on the 2/30/180 rule and/or not being prescribed an antidepressant.

Patient demographics		Analysis cohort	Received therapy[1]	PHQ-measured[2]	PGx data collected[3]	Excluded population
Age*	Mean ± SD (IQR range)	53.8 ± 19.5 (39, 70)	57.1 ± 18.8 (43, 71)	46.7 ± 22.5 (24, 67)	53.7 ± 19.7 (40, 69)	54.9 ± 16.9 (37, 68)
Gender	Male	14,007 (25.8%)	3,021 (24.1%)	1,072 (29.8%)	181 (21.6%)	35,665 (27.9%)
	Female	40,250 (74.2%)	9,517 (75.9%)	2,654 (71.2%)	656 (78.4%)	91,998 (72.1%)
	Unknown/ Unspecified	2 (~0.0%)	1 (0.0%)	0 (0.0%)	0 (0.0%)	9 (~0.0%)
Race	White	50,084 (92.3%)	11,119 (88.7%)	3,123 (83.8%)	736 (87.9%)	116,252 (91.1%)
	Black	2,815 (5.2%)	974 (7.8%)	437 (11.7%)	68 (8.1%)	6,913 (5.4%)
	Asian Pacific Islander	527 (1.0%)	213 (1.7%)	63 (1.7%)	11 (1.3%)	1,047 (0.8%)
	American Indian	64 (0.1%)	26 (0.2%)	6 (0.2%)	1 (0.1%)	152 (0.1%)
	Alaskan Native	4 (0.0%)	1 (0.0%)	0 (0.0%)	0 (0.0%)	14 (0.0%)
	Unknown/Unspecified/ Missing/Declined	755 (1.4%)	206 (1.6%)	97 (2.6%)	21 (2.5%)	3,294 (2.6%)
Ethnicity	Not Hispanic or Latino	51,412 (94.8%)	11,809 (94.2%)	3,501 (94%)	786 (93.9%)	119,924 (93.9%)
	Hispanic or Latino	493 (0.9%)	146 (1.2%)	50 (1.3%)	6 (0.7%)	1,156 (0.9%)
	Unspecified/Declined	2,354 (4.3%)	584 (4.7%)	175 (4.7%)	45 (5.4%)	6,592 (5.2%)
Vital status	Presumed Alive	52,701 (97.1%)	12,241 (97.6%)	3,618 (97.1%)	833 (99.5%)	116,821 (91.5%)
	Known Deceased	1,558 (2.9%)	298 (2.4%)	108 (2.9%)	4 (0.5%)	10,851 (8.5%)
Total		54,259 (100%)	12,539 (23.1%)	3,726 (6.9%)	837 (1.5%)	126,663

[1:]Patients that have received psychotherapy
[2:]Patients that have PHQ scores within their EHR
[3:]Patients that have PGx data recorded.

Within the 4,049 (2.22%) patients that do not have a race associated with their EHR data, there are 1,128 (0.62%) subjects with an omission for race (Table 3). 1,189 (0.65%) subjects "Declined" to specify their race. 1,714 (0.94%) subjects were "Not Specified". Eighteen

(0.0099%) of individuals were of "Unknown" race. A total of 4,049 (2.22%) subjects have missing race. Within the 8,946 (4.94%) patients that do not have an ethnicity associated with their EHR data, there are 2,961 (1.62%) subjects with an omission for ethnicity. 2,773 (1.52%) subjects "Declined" to specify their ethnicity. 3,212 (1.77%) subjects were "Not Specified" for ethnicity. There were 2,060 individuals that had both missing ethnicity and race. 836 patients were "Not Specified" for both ethnicity and race.

3.3.1.2 Diagnoses

Of the diagnosis codes, most (64.2%) are ICD-10 codes, while 35.8% are ICD-9 codes. Most (n = 6) fields in the diagnoses table did not have any missing values, except for the "Diagnosis to date" field, which was entirely missing and therefore does not seem to be used by clinicians in the EHR. There were 31,558 unique diagnosis codes. The most frequent diagnoses were need for prophylactic vaccination and inoculation against influenza, anxiety, depressive disorder not elsewhere classified, and depression (Table 4).

Table 4 Top 20 Most Frequent Diagnoses.

Diagnosis name	Number of patients (%)
Need for prophylactic vaccination and inoculation against influenza	81,860 (45.00%)
Anxiety	53,173 (29.23%)
Depressive disorder, not elsewhere classified	53,078 (29.17%)
Depression	49,472 (27.19%)
Essential hypertension	49,250 (27.07%)
Vitamin D deficiency	38,406 (21.11%)
Other and unspecified hyperlipidemia	34,536 (18.98%)
Unspecified essential hypertension	30,148 (16.57%)
Mixed hyperlipidemia	28,529 (15.68%)
Anxiety and depression	25,901 (14.24%)
Gastroesophageal reflux disease without esophagitis	25,330 (13.92%)
Anxiety state, unspecified	25,149 (13.82%)
Pure hypercholesterolemia	15,417 (8.47%)
Unspecified hypothyroidism	15,097 (8.30%)
Type II or unspecified type diabetes mellitus without mention of complication, not stated as uncontrolled	13,682 (7.52%)
Acquired hypothyroidism	13,150 (7.23%)
Essential hypertension, benign	12,430 (6.83%)
Atrial fibrillation (HCC)	4,892 (2.69%)
Long term (current) use of anticoagulants	4,667 (2.57%)
Long term current use of anticoagulant therapy	3,322 (1.83%)

3.3.1.3 Encounters

There were 176 types of encounters. The most prevalent encounter types were “Appointment” (N=5,777,297; 15.41%), Telephone (N=5,103,610; 13.61%), History (N=4,203,443; 11.21%), Office Visit (N=4,124,605; 11.0%), Refill (N=3486344; 9.3%), and Scan (3277836; 8.74%) (Table 5). Patients were treated at a total of 5,202 locations. Most encounters (N=5,399,311; 14.4%) occurred at an “External Department” (Table 6).

Table 5 Top 14 Most Frequent Encounter Types.

Encounter type	Number of encounters (%)
Appointment	5,777,297 (15.41%)
Telephone	5,103,610 (13.61%)
History	4,203,443 (11.21%)
Office visit	4,124,605 (11%)
Refill	3,486,344 (9.3%)
Scan	3,277,836 (8.74%)
Lab results	1,699,872 (4.53%)
Patient email	1,282,627 (3.42%)
Imaging	1,058,141 (2.82%)
Hospital encounter	550,954 (1.47%)
ER report	445,220 (1.19%)
Letter (Out)	401,157 (1.07%)
Nurse visit	364,585 (0.97%)
Informational	360,354 (0.96%)

There were 9,616 encounters with no encounter type. There were 155 encounters that were of "0" encounter type, and 19,336 encounters that were of "79" encounter type. Therefore, there were 29,107 (0.08%) missing encounter types. Department facility ID and location had relatively low missingness at both 0.42%. Appointment status had 69.88% missingness.

Table 6 Top 20 Encounter Locations

Location	Number of Encounters at Location (%)
External department	5,399,311 (14.4%)
UPMC General Internal Medicine, Oakland (Pittsburgh, PA)	450,285 (1.2%)
Mainline Medical Associates (Altoona, PA)	306,489 (0.82%)
Primary Care Partners (Fairview, PA)	225,491 (0.6%)
Summit Family Practice (Erie, PA)	224,422 (0.6%)
Northern Medical Associates (Wexford, PA)	208,928 (0.56%)
Renaissance Family Practice (Verona, PA)	208,289 (0.56%)
Healthy Families Primary Care (Erie, PA)	202,959 (0.54%)
West Erie Medical Group (Erie, PA)	199,817 (0.53%)
John Chantz and Associates (Pittsburgh, PA)	195,107 (0.52%)
Health Center Associates (Pittsburgh, PA)	189,762 (0.51%)

3.3.1.4 Medication Prescription Events

There were 181,483 (99.7%) individuals with medication prescription events, and 12,779,542 medication prescription events in total. Of the 12,779,542 prescription events, 7.25% of prescriptions were for SSRIs, 2.11% were for SNRIs and 0.62% were for TCAs. There were 27,947 unique medications prescribed, which was represented by 5,671 generic medication names. The top generic medications ordered were levothyroxine, sertraline, and hydrocodone/acetaminophen (Table 7).

Table 7 Top 20 generic medications ordered.

Generic medication name	Number of orders (%)
Levothyroxine	330,617 (2.59%)
Sertraline	287,060 (2.25%)
Hydrocodone/acetaminophen	273,745 (2.14%)
Citalopram	230,748 (1.81%)
Lisinopril	215,027 (1.68%)
Omeprazole	208,676 (1.63%)
Atorvastatin	206,119 (1.61%)
Lorazepam	186,034 (1.46%)
Bupropion	180,070 (1.41%)
Fluoxetine	175,151 (1.37%)
Alprazolam	173,854 (1.36%)
Metformin	170,273 (1.33%)
Escitalopram	168,416 (1.32%)
Gabapentin	165,813 (1.30%)
Prednisone	159,430 (1.25%)
Venlafaxine	154,318 (1.21%)
Simvastatin	148,815 (1.16%)
Amlodipine	144,266 (1.13%)
Fluticasone	136,884 (1.07%)
Albuterol	130,701 (1.02%)

The medication order table had low missingness for order date, medication order ID, and medication name (Figure 2). However, there were 123,979 medications (722 types of medications) prescribed that did not have an associated generic medication name. Almost half (48.39%) of

medications prescribed did not have an associated dose or medication unit. About one fifth (21.24% and 20.25%, respectively) of medications prescribed did not have a quantity specified, or a refill specified.

There were 2,713,680 (21.23%) orders with no quantity of medication specified (Figure 2). In addition, there were 4 orders with a quantity of “0”, 5 with “0 capsule”, 3 with “0 g”, 58 with “0 Tab”, 256 with “0 tablet”. The data field for “Instructions” had too much variation to include in the analysis. Namely, there were 1,178,139 unique instructions out of 12,263,894 total instructions, with 515,644 orders having no instructions, and 4 orders having the instruction “0”.

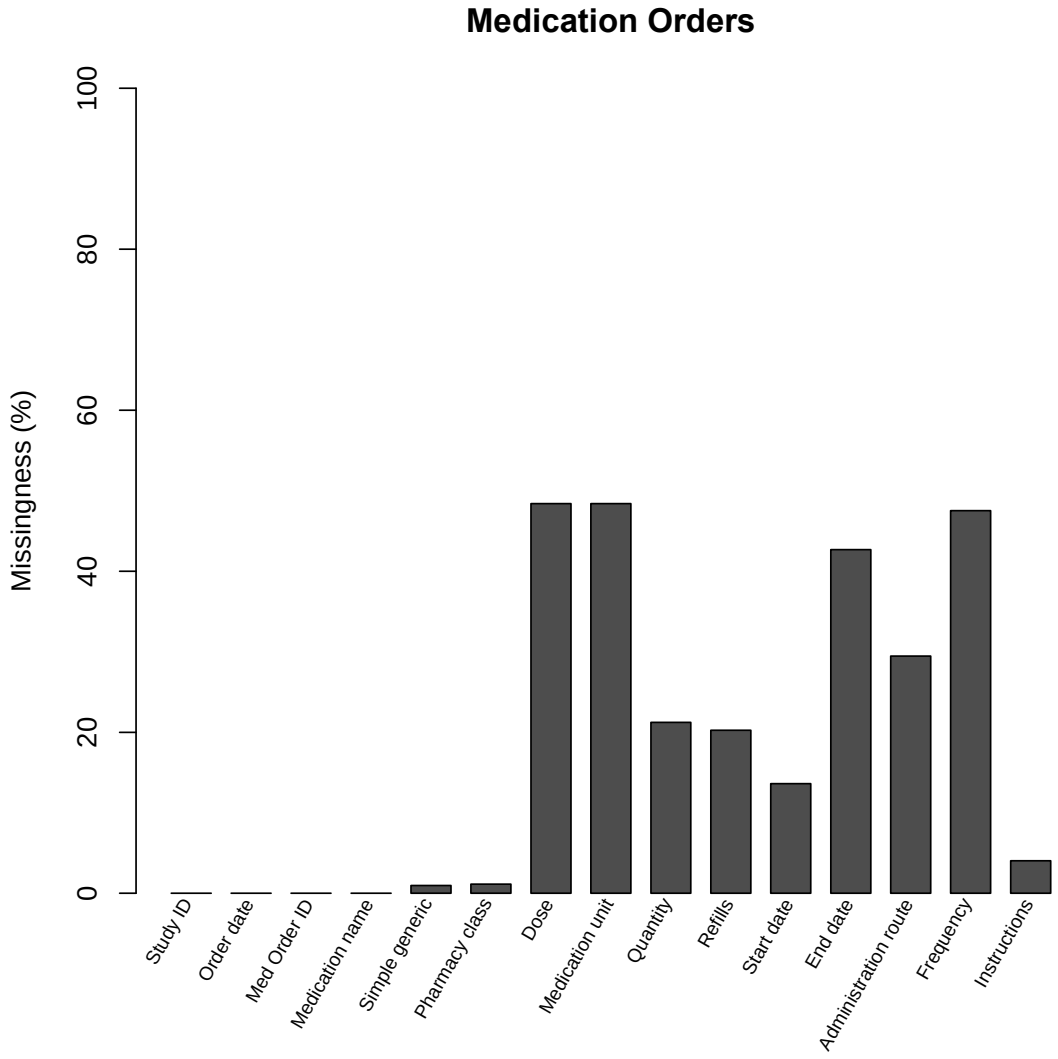


Figure 2 Medication Order Variable Missingness

3.3.1.5 Medication Fills

There were 166,395 (91.5%) patients that filled prescriptions, leaving nine percent of patients that did not fill their prescriptions. In examining medication fills, there were 2,930 unique generic medications filled within the EHR database. For individuals that filled prescriptions, patients filled an average of 348.7 prescriptions.

Medication fills had relatively low missingness for fill date, drug name, simple generic drug name, drug code, and amount (Figure 3). However, almost half (47.67%) of medication fills were missing an associated medication unit. In addition, almost half (46.46%) of medication fills were missing a medication unit strength.

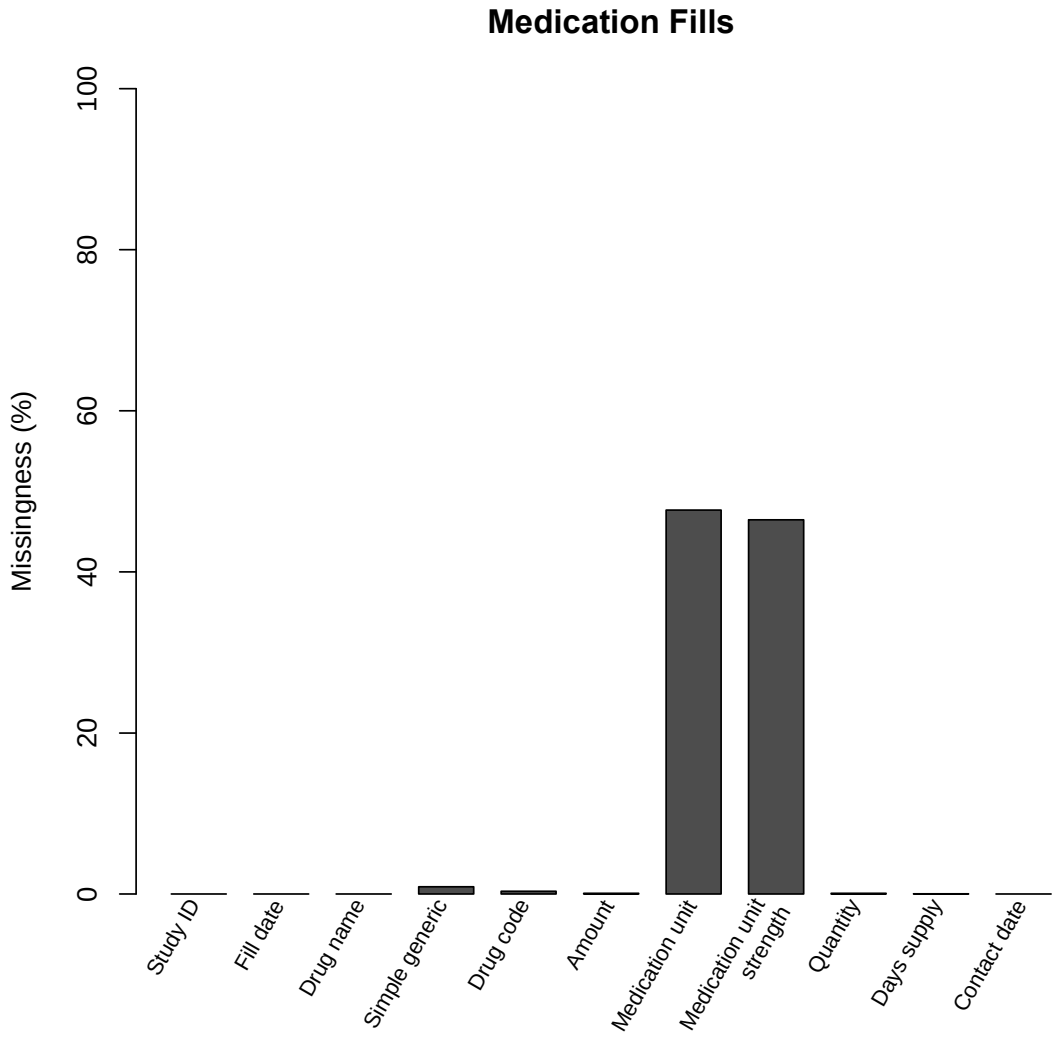


Figure 3 Medication Fill Variable Missingness

The analysis of medication fills and orders revealed the necessity of using regular expressions to standardize the drug name field, given that many medications start with symbols

(e.g., "-" and "^") in the EHR. Other inconsistencies were revealed in the drug code field, where 5,134 fills had a drug code of zero and 200,353 had no drug code, for a total of 205,487 missing drug codes. There were 43,715 fills of zero amount, and none of the instances of fills were missing for the "Amount" field.

For "Med_Unit", there were 5,533 orders with missing data, and 27,653,091 that were "Not Specified". For "Med_Unit_Strength" there were 623,599 fills with a strength of zero, and 26,335,344 fills with no specified strength. There were 43,715 fills of zero quantity. There were 88 fills that had zero days of supply.

3.3.1.6 Depression Questionnaire Data

Five percent (N=10,226) of patients had depression severity questionnaire data on file. Of patients that had questionnaire data within the EHR, the average number of questionnaires taken for each patient was 25.8 questionnaires. The median number of questionnaires taken for each patient was 18. There were 264,120 total patient questionnaires within the EHR database. Forty two percent of patients (N=4,362) had one day of questionnaire data (Table 8). Most patients with depression severity questionnaire data have responses to the PHQ-4 questionnaire (Table 9). There were 14 instances where a patient had multiple questionnaires taken on the same day.

Table 8 Distribution of number of days of questionnaire data.

Days of Questionnaire Data	Frequency
1	4,362 (42.6%)
2	2,818 (27.6%)
3	1,350 (13.2%)
4	697 (6.8%)
5	420 (4.1%)
6	247 (2.4%)
7	139 (1.6%)
8	75 (0.7%)
>8	118 (1.1%)
Sum	10,226

Table 9 Distribution of depression questionnaire data for patients within dataset.

Form ID	Form Name	N
16000494	PHQ-4	188,770
1400005303	PHQ 2-8	42,547
16000600	PHQ-2	30,823
16000587	PHQ-8 (FULL QUESTIONNAIRE)	1,800
1400005300	UPMC HEALTHTRAK PHQ9 QUESTIONNAIRE	178
1400005304	UPMC HEALTHTRAK PHQ-8 QUESTIONNAIRE 2	8
16000490	BRANCH FROM PHQ-4 TO PHQ-8	8
Sum		264,134

For patients that had questionnaire data, the fields were highly populated with no missingness besides the "Answer" field, which had 19.63% missingness. PHQ-8 scores were mapped over time for individual patients (Figure 4). According to group means and median PHQ scores, the majority of patients were classified to have moderate to moderately severe depression (Table 10). A one-way ANOVA revealed a statistically significant difference in PHQ scores for gender ($P = 0.011$) and for gender and race together ($P = 0.013$), but not for race alone ($P = 0.306$). A Tukey HSD for multiple comparison of means for gender and race found White women had significantly higher PHQ scores than White men ($P = 0.024$), however no other gender and race

combinations were significantly different ($P > 0.050$) (Figure 5). Shapiro-Wilks test for normality of PHQ score distributions for White men ($P = 0.0005$), White women ($P < 0.0001$), and Black women ($P = 0.0001$) failed to reject the null hypothesis that the distributions were normal, while the Shapiro-Wilks test on PHQ score distributions for Black men ($P = 0.06$), Asian women ($P = 0.21$), and Asian men ($P = 0.29$) rejected the null hypothesis (Figure 6).

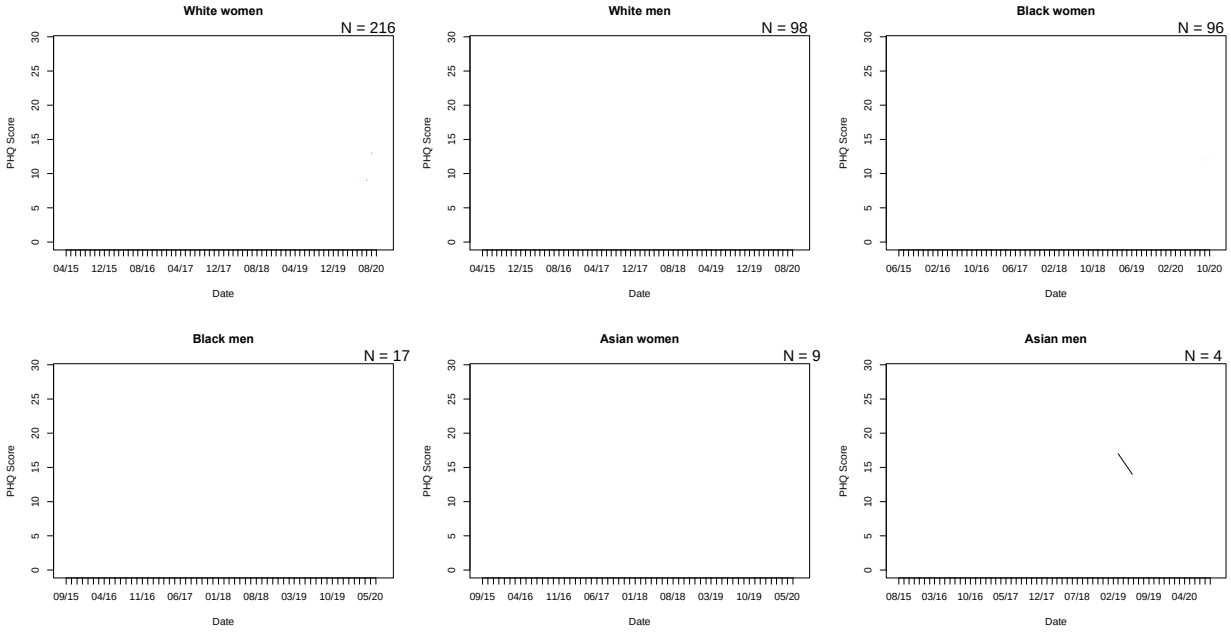


Figure 4 PHQ scores over time for MDD patients, where each color and line variation represents an individual patient within their reported race and gender subgroup.

Table 10 Summary statistics of PHQ scores subset by patients' race and gender.

Race	Gender	Number of patients with at least one PHQ score	Number of patients with > 1 PHQ scores	Mean PHQ score (Mode)	Median PHQ score (Range)
White	Female	627	216	14 (13)	14 (1, 24)
	Male	272	98	13 (12)	13 (2, 24)
Black	Female	184	96	14 (16)	14 (2, 24)
	Male	36	17	15 (13)	15 (2, 24)
Asian Pacific Islander	Female	24	9	14 (15)	15 (4, 22)
	Male	8	4	13 (15)	14 (7, 20)

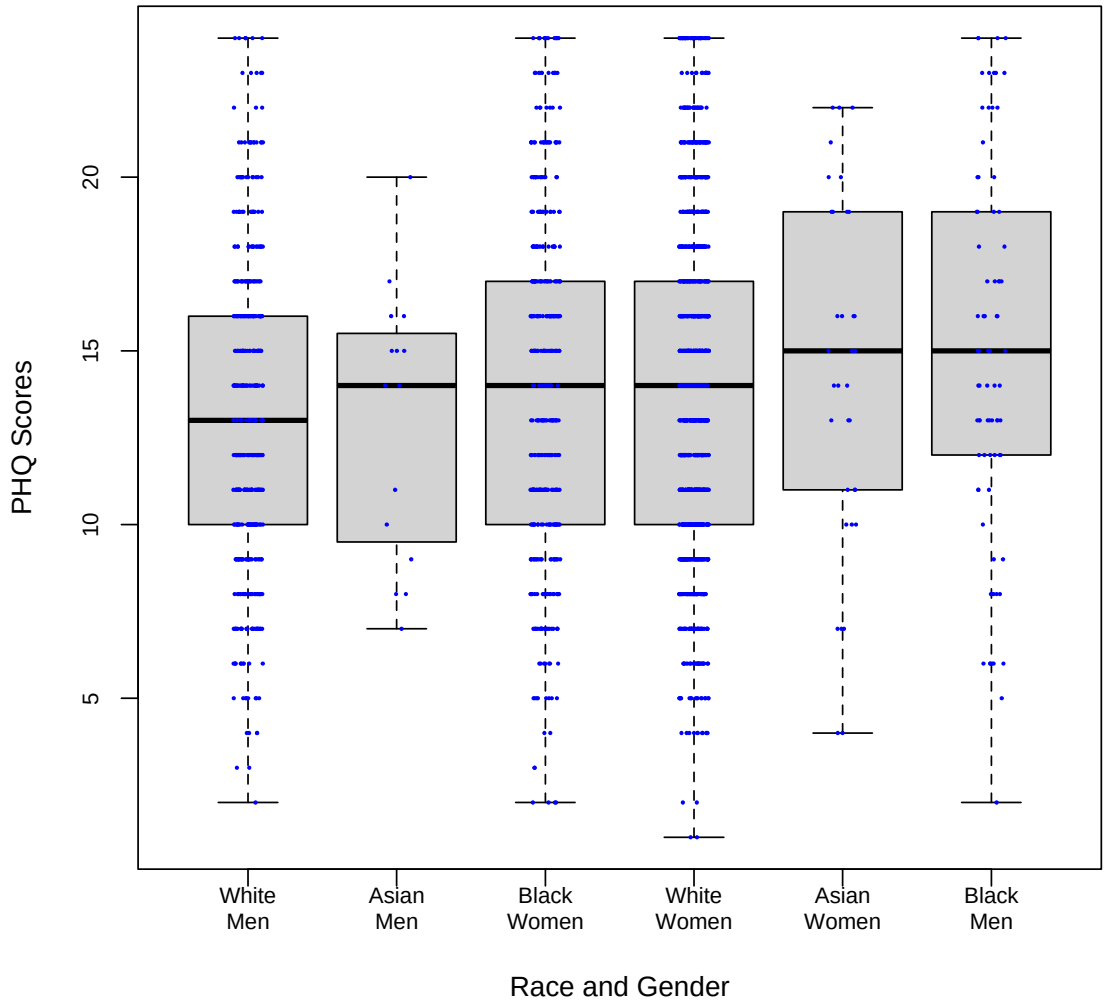


Figure 5 Boxplots of PHQ scores for patients subset by race and gender.

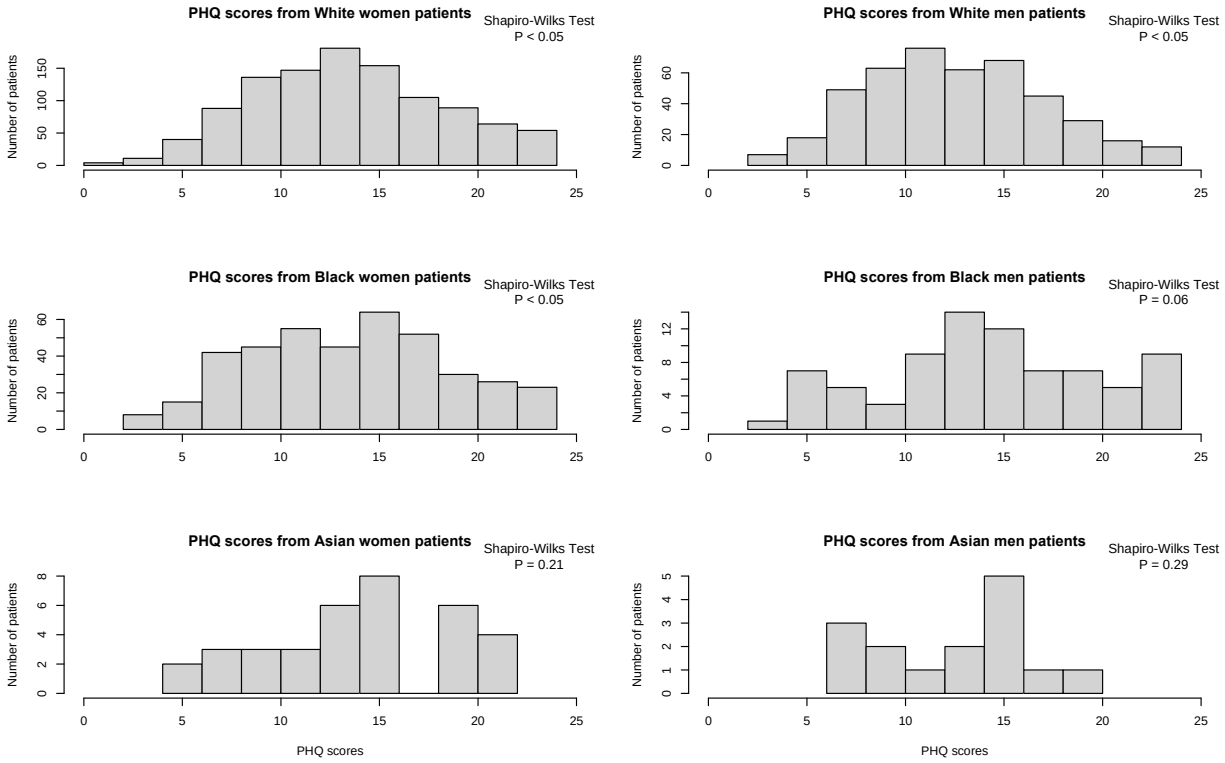


Figure 6 Histograms of patients' PHQ score distributions subset by race and gender.

Mean PHQ score slopes were lowest for Asian Pacific Islander women, and highest for white women (Table 11). The average PHQ score slopes across all patients was 0.016. A negative slope reflected a decrease in depression severity, while a positive slope reflected an increase in depression severity. One-way ANOVA comparing PHQ score slope means revealed no statistically significant difference when comparing race (P = 0.382), gender (P = 0.586), or race and gender together (P = 0.586) (Figure 7).

Table 11 Summary statistics of PHQ scores slopes subset by patients' race and gender.

Race	Gender	Mean PHQ score slope	Median PHQ score slope	Minimum PHQ score slope	Maximum PHQ score slope
White	Female	0.016	0.002	-0.229	1.000
	Male	0.005	0.000	-0.143	0.126
Black	Female	0.005	0.001	-0.375	0.220
	Male	0.006	0.000	-0.022	0.067
Asian Pacific Islander	Female	-0.027	0.013	-0.387	0.048
	Male	0.012	0.006	0.000	0.036

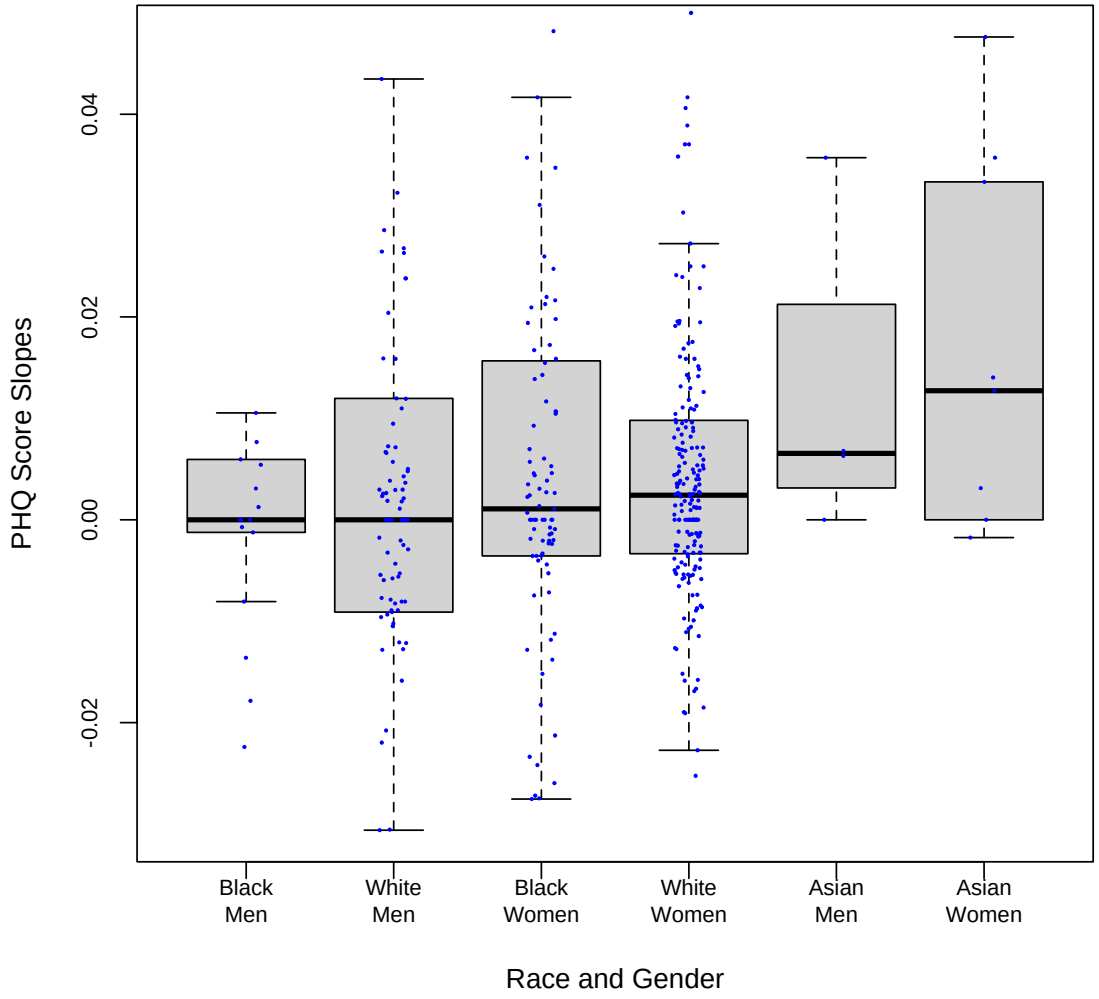


Figure 7 Boxplots of PHQ score slopes for patients subset by race and gender.

3.3.1.7 Vitals

Most patients (97.8%) had vitals data on file. Height had a missingness of 22.94%. Both the average and median height was 5'4". There were 712,563 (15.48%) vital assessments with no weight taken, and 5 assessments with a zero for weight. The average weight of individuals in the EHR database was 190.1 pounds, and the median was 183.05 pounds. BMI had 16.22% missingness. The average BMI was 32.19, which is considered obese. The median BMI was 30, which is also considered obese. Underweight individuals made up 1.1% (N=602) of the EHR dataset. Healthy individuals made up 20.8% (N=11,346) of the EHR dataset, while overweight individuals and obese individuals made up 27.7% (N=15,115) and 49.6% (N=27,050), respectively.

There were 60 assessments with a systolic blood pressure of 0, and 755,449 assessments with "NA" for systolic blood pressure, for a total of 16.4% missingness. There were 986 assessments with a diastolic blood pressure of 0, and 755,447 with NA for diastolic blood pressure, for a total of 16.4% missingness. There was 29.29% missingness for blood pressure position. For pulse, there were 1,563,326 assessments with an NA or zero for pulse, contributing to a total missingness of 33.97%. Temperature had 60.16% missingness. There were 26 assessments with a respiratory rate of zero, and 3,384,311 assessments had an NA for respiratory rate, for a total of 73.53% missingness.

3.3.1.8 Missingness Summary

Due to the fact that missingness was low for variables necessary to model creation (demographic data, diagnosis codes, medication orders, vitals data) (Table 3), imputation was determined to be unnecessary because imputation with 5% missingness has been shown to have

negligible benefit [102,109]. Most of the missingness came from the medication fills and orders tables, specifically from the dose information, start date and end date, quantity, instructions, etc. which were not used in our models. Dose information was captured in the medication name for orders. Therefore, a complete case analysis was determined as most appropriate for the study.

3.4 Discussion

Perhaps atypical for EHR data, missingness was low for the clinical variables of interest to construct models. The demographic distributions of the cohorts were expected, in that the majority of the depression analysis cohort were females. Also, the demographic breakdown of the cohorts in terms of race and ethnicity were also typical of western Pennsylvania. However, in order for this pipeline to be generalizable to a diverse set of patients, samples from a wider range of race and ethnicities must be sought out in order for a more equitable and fair pipeline [110,111]. The lack of diversity in study populations is a common misfortune of clinical research today, and steps must be taken to overcome this major limitation [112]. Future directions could involve running this analysis pipeline on a hospital's EHR that cares for a more diverse patient population.

Depression severity score data is a rich quantitative measure of patients' symptoms. PHQ scores over time were mapped, however future directions could involve subsetting these PHQ scores to uncover patterns based on clinical features like diagnosis codes or treatment response. In addition, as EHR data analyses expand, and in turn, PHQ data for patients are examined more routinely, it might prove to be beneficial to expand the collection of this quantitative data describing patients' symptoms as opposed to relying on whether there was a dose change or drug switch to proxy treatment response.

Most patients for which there were PHQ scores within the EHR were classified as having moderate to moderately severe depression according to PHQ score thresholds. This could reflect a bias that most patients had PHQ scores documented when cases were more severe as opposed to more mild cases that did not appear to be captured often. Future directions could involve investigating whether these PHQ score distributions are an accurate reflection of patients' depression severities, or whether there are more mild depression cases that are not captured or documented in the EHR.

There was no statistically significant difference in mean PHQ score slopes when comparing subsets of patients based on race, gender, or race and gender together. The average PHQ score slope was almost zero (0.016), which reflects a lack of change in depression severity for patients over time documenting PHQ scores in the EHR. There was also no statistically significant difference in PHQ score means when comparing subsets of patients based on race, however there was a significant difference when comparing mean PHQ scores for patients according to subsets of gender and both gender and race together. When comparing gender and race together, only White men and White women had a statistically significant difference in PHQ score means. However, there were more PHQ scores documented within the EHR for White women, White men, and Black women. This was reflected in the results of the Shapiro-Wilks tests for normality failing to reject the null hypothesis of normal distribution of scores for White women, White men, and Black women, and rejecting the null hypothesis for race and gender subsets for which there were lower sample sizes: Black men (N=36), Asian women (N=24), and Asian men (N=8). Future directions could be dedicated towards collecting PHQ scores from populations that are not White in order to re-run these analyses with enough samples to reflect a normal distribution of PHQ

scores. Future directions should also uncover the factors contributing to the racial disparities in PHQ score documentation in the EHR.

4.0 Pharmacogenomics Variant Translation

4.1 Introduction

Pharmacogenomics variant translation can be performed through software tools, such as PharmCAT [113]. PharmCAT takes variant call format (VCF) data from sequencing and genotyping technologies and assigns diplotypes (one haplotypes for each chromosome) using established translation tables [114]. These diplotypes can also be the star-allele definition, which are haplotypes that are agreed upon in the field of PGx and drive prediction of a phenotype. From the assigned diplotypes, the associated the diplotype or star allele can be associated with a CPIC guideline recommendation for prescribing, if the diplotype has an associated guideline.

There has been work done to establish reproducible pharmacogenomic pipelines due to the necessity of standardization in translation PGx to clinical care systematically, robustly, and flexibly [115]. We will build off of other software PGx translation tools and reproducible pipelines to implement our own PGx pipeline, built off of the workflow established in the Empey Lab.

4.2 Methods

4.2.1 Pharmacogenomic Variant Translation

A pipeline was constructed to annotate pharmacogenomics data for patients within the Pitt + Me Discovery cohort based on an established workflow of variant calling in the Empey Lab. PGx data were standardized and normalized. PGx haplotype variant data was annotated according to standardized phenotypes from CPIC definitions. Standardization took place using translation tables that label star alleles.

An example variant call format (VCF) file in 4.2 format was used to demonstrate proof of concept as additional patients are enrolled in the Pitt + Me Discovery cohort and therefore additional data is contributed to the analysis pipeline. There are two genes of particular interest to MDD based on CPIC guidelines: CYP2C19 and CYP2D6. Using a translation table for CYP2C19 and a table for CYP2D6, patients were annotated according to the appropriate star allele based on their haplotype. For example, a patient that has a thymine allele at rs12248560, a guanine at rs3758581would be classified as CYP2C19*17. Whereas the patient would be a CYP2C19*18 if they had an adenine at rs138142612. The alleles are conveyed through a genotype column containing “0” for the reference allele, a “1” for the first allele noted as the alternate allele, or a “2” for the second allele noted as the alternate allele. Diploid calls are connoted through a “|” separating the two values in the genotype column. Haploid calls are expressed through only one allele value. Patients would be annotated based on a finite number of well-established phenotype definitions (ultra-rapid metabolizer, extensive/normal metabolizer, intermediate metabolizer, or poor metabolizer) that would serve as input features into the statistical models.

4.3 Results

Patients are still being enrolled in the Pitt + Me Discovery cohort, and pharmacogenomics data is still being collected. Currently, there are 42 samples collected, which is too low of a sample size to be included for these statistical models.

An example patient, “REI70A170” was annotated to have a CYP2C19*1/*1 genotype because no variants were found to be different from the reference sequence among alleles tested. Another example patient was translated to be a CYP2C19*2/*2, which has a splicing defect that results in a truncated protein and therefore has no function. The first patient was an extensive (normal) metabolizer, while the second patient was a poor metabolizer.

4.4 Discussion

The CYP2C19*2/*2 patient that was a poor metabolizer would likely not experience treatment response or symptom remission because of the splicing defect leading to a protein with no function, and therefore the inability to metabolize SSRIs. For these cases of patients PGx data would be instrumental to understanding ahead of the therapeutic wait time of eight to twelve weeks that first-line therapeutics will not be metabolized by this patient and therefore the patient will not experience a treatment response or symptom remission. Studies have examined the prevalence of patients that are poor metabolizers not responding to first-line therapies due to a poor metabolizer phenotype [116–118].

Future directions include expanding the pharmacogenomic variant translation and incorporating labeled pharmacogenomics phenotype information from Pitt + Me Discovery cohort

patients diagnosed with MDD in downstream models to aid in treatment response and symptom remission prediction. As additional patients are enrolled and pharmacogenomics panel data is collected, the reproducible pipeline will be re-run, metabolizer phenotypes will be called, and the phenotypes will be used as additional variables to inform statistical models.

5.0 Electronic Phenotyping

5.1 Introduction

Phenotyping, the identification of patients with similar outcomes or conditions, is important to consistently perform cohort analyses that can ascertain valuable information about accurately grouped individuals. Phenotyping is paramount to clinical research, especially as it relates to translating research to the clinic, comparing drugs/treatments, and clinical decision support [119]. However, phenotyping in the EHR, or electronic phenotyping, is no simple task due to the incomplete, biased, heterogenous, and dynamic nature of EHRs [120]. It also should be emphasized contents of the EHR can be incorrect or recorded in error [121].

The Electronic Medical Records and Genomics (eMERGE) consortium of nine academic medical centers have worked to establish EHR-generated phenotyping algorithms to conduct repeatable and accurate cohort studies [122–124]. Hripcsak and Albers (2013) note the challenges of using EHRs to represent the patient's true state, and how the recording process of data into the EHR creates a deportation from the patient's true state [125].

However, there have been major advances in electronic phenotyping that allow researchers to traverse the challenges presented by EHRs in the heterogeneity between patient EHR data, multivariate data types, and missingness present in the EHR. Namely, researchers implement rule-based methods, natural language processing (NLP), machine learning frameworks, and combinatorial approaches to conduct electronic phenotyping. The review by Banda et al. (2018) found that 19 papers used rule-based methods, 35 papers used NLP, 25 used machine learning, and 10 used combinatorial approaches to conduct electronic phenotyping [119].

Rule-based methods are a traditional methodology for conducting electronic phenotyping, and use structured data fields like diagnosis codes, medications, procedures, and lab codes to define the inclusion and exclusion criteria for cohorts. Rule-based methods work well for outcomes or conditions that are defined in the EHR with clear structured data elements. Rule-based methods that implement multiple structured data elements, for example both diagnosis codes and medications, demonstrate increased performance [126,127]. eMERGE has published rule-based phenotypes that can be accessed on phekb.org [128].

Other methods for electronic phenotyping include text mining from unstructured data like clinical notes and using NLP to ascertain meaning and phenotypic information. In addition, studies also implement machine learning to conduct electronic phenotyping. In this study, we will use the eMERGE rule-based electronic phenotype definition for depression.

5.2 Methods

5.2.1 eMERGE Electronic Phenotype Implementation

The PheKB electronic phenotype for MDD was applied to the original cohort of 181,930 patients. This electronic phenotype is made up of inclusion ICD-9/10 codes and exclusion ICD-9/10 codes, along with a temporal qualification that patients must have evidence of depression on 2 calendar days, at least 30 days apart, and no greater than 180 days apart.

5.2.2 Medication List Construction

In collaboration with Dr. Tanya Fabian (Associate Professor of Pharmacy and Therapeutic and Psychiatry) an antidepressant list was constructed and edited to accurately reflect clinical practice. The antidepressant list was then labelled according to drug class and function as independent therapeutic agents or as augmenting agents prescribed in combination with independent therapeutic agents.

5.2.3 Electronic Phenotyping for Outcome Classification

Once the electronic phenotype for MDD was applied to allow for a final dataset of patients, additional electronic phenotypes were implemented to allow for greater granularity in describing patients' experience with depression. Patients prescribed zero or one antidepressant were characterized as having "early depression", whereas patients that failed two or more adequate antidepressant trials are considered to experience "treatment-resistant depression". Patients' first PHQ scores within the EHR database were also classified for whether they had mild, moderate, moderately severe, or severe depression (Table 12). Further, depression severity was classified as patients progressed through their antidepressant treatment sequence. Subsequent encounters from the original diagnosis encounter were labeled based on PHQ score thresholds to judge symptom remission and changes in PHQ scores were used to ascertain treatment response (Table 13). Initial antidepressant therapies were used in the statistical models so initial antidepressants were treated as separate, independent exposures. Initial and subsequent antidepressant therapies were included in the Markov chain models also took the previous antidepressant therapy into consideration when

calculating transition probabilities between antidepressants. Therefore, the Markov chain models did not treat antidepressants as separate, independent exposures.

Table 12 PHQ scores and descriptions

PHQ-9 Score	Depression severity	Proposed treatment actions
0-4	None-minimal	None
5-9	Mild	Watchful waiting; repeat PHQ-9 at follow-up
10-14	Moderate	Treatment plan, consider counseling, follow-up and/or pharmacotherapy
15-19	Moderately severe	Active treatment with pharmacotherapy and/or psychotherapy
20-27	Severe	Immediate initiation of pharmacotherapy and, if severe impairment or poor response to therapy, expedited referral to a mental health specialist for psychotherapy and/or collaborative management

Table 13 Outcome definitions parallel to STAR*D outcome definitions (PHQ instead of QIDS-C_{16} scores)

Outcome	Outcome definition	PHQ score
Remission	Remitter	<=4
	Probable remitter	5-9
	Non-remitter	>=10
Response	Responder	>=50% reduction compared to baseline
	Probable responder	40-50% reduction compared to baseline
	Non-responder	< 40% reduction compared to baseline

Once patient encounters were labelled for both remission and response based on PHQ scores, individual antidepressant treatments will be labeled for whether the treatment was successful or not. A "successful antidepressant" is one for which the patient experiences symptom remission or treatment response (Table 13) at the subsequent encounter as demonstrated by the PHQ score. In addition, to ensure an adequate medication trial, the antidepressant must continue to be prescribed for at least eight weeks at subsequent encounters. This electronic phenotype description will likely be specific enough to generate high-quality and high-fidelity data for successful antidepressants based on similar studies like STAR*D. However, in the case that these definitions are unable to confirm successful antidepressants for patients, we can employ unsupervised or supervised learning to generate additional features associated with successful antidepressants.

An “unsuccessful antidepressant” was defined as one for which there is a change in antidepressant in the subsequent encounter, and/or the PHQ score deems no symptom remission or treatment response. Similar to an “unsuccessful antidepressant”, an outcome of “toxicity” was labelled for when there was a discontinuation or dose decrease of an antidepressant in the subsequent encounter.

5.3 Results

5.3.1 eMERGE Electronic Phenotype Implementation

As expected, zero patients in the original patient cohort of 181,931 had an ICD-9 or ICD-10 code from the exclusion criteria. There were 23 ICD-9/-10 codes for MDD present in the total EHR database, where the majority of patients in both the total EHR database and the analysis cohort had the F32 ICD-10 code (51.86% and 76.56%, respectively) (Table 14). MDD patients in the total EHR database had an average of nine MDD ICD-9/-10 codes over the course of their EHR data. The number of MDD ICD-9/-10 codes over the course of patients’ EHR data ranged from two to 1,366 MDD ICD-9/-10 codes, where the median was five codes (3.91%, N=7,127 patients) and the mode was two codes (11.35%, N=20,651 patients).

Within the analysis cohort, the average number of MDD ICD-9/-10 codes was 12 codes, with the range being from three to 1,366 codes. The median number of MDD ICD-9/-10 codes was seven (7.29%, N=3,956 patients), and the mode was three (12.07%, N=6,549 patients).

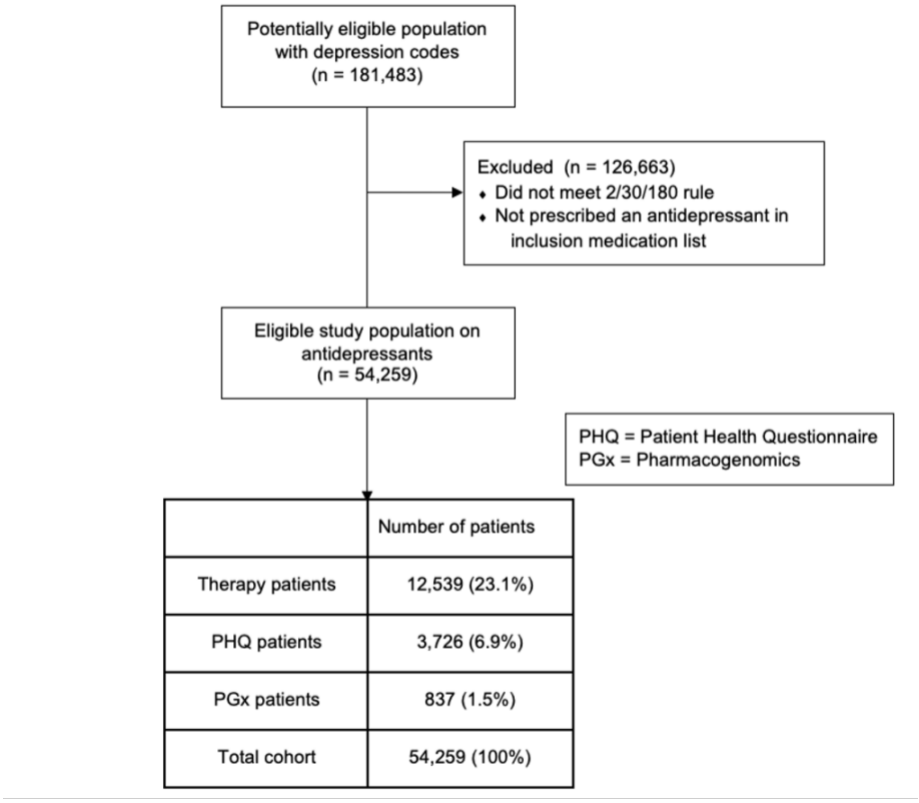
Table 14 Distribution of MDD ICD-9/-10 codes across patients.

ICD-9 or ICD-10	Code	Description	Number of patients with code in total EHR database (N=181,931)	Number of patients with code in analysis cohort (N=54,259)
ICD-10	F32.	Major depressive disorder, single episode	94,349 (51.86%)	41,543 (76.56%)
	F33.	Major depressive disorder, recurrent	36,334 (19.97%)	22,314 (41.12%)
	F34.	Persistent mood [affective] disorders	6,983 (3.84%)	4,046 (7.46%)
	F39.	unspecified mood [affective] mixed	5,914 (3.25%)	2,494 (4.60%)
	F43.21	Adjustment disorder with depressed mood	6,942 (3.82%)	3,198 (5.89%)
ICD-9	296.20	Major depressive affective disorder, single episode, unspecified	6,813 (3.74%)	4,805 (8.86%)
	296.21	Major depressive affective disorder, single episode, mild	6,813 (3.74%)	4,805 (8.86%)
	296.22	Major depressive affective disorder, single episode, moderate	6,813 (3.74%)	4,805 (8.86%)
	296.23	Major depressive affective disorder, single episode, severe	6,813 (3.74%)	4,805 (8.86%)
	296.25	Major depressive affective disorder, single episode, in partial or unspecified remission	6,813 (3.74%)	4,805 (8.86%)
	296.26	Major depressive affective disorder, single episode, in full remission	6,813 (3.74%)	4,805 (8.86%)
	296.30	Major depressive affective disorder, recurrent episode, unspecified	3,128 (1.72%)	2,156 (3.97%)
	296.31	Major depressive affective disorder, recurrent episode, mild	3,128 (1.72%)	2,156 (3.97%)
	296.32	Major depressive affective disorder, recurrent episode, moderate	3,128 (1.72%)	2,156 (3.97%)
	296.33	Major depressive affective disorder, recurrent episode, severe, without mention of psychotic behavior	3,128 (1.72%)	2,156 (3.97%)
	296.35	Major depressive affective disorder, recurrent episode, in partial or unspecified remission	3,128 (1.72%)	2,156 (3.97%)
	296.36	Major depressive affective disorder, recurrent episode, in full remission	3,128 (1.72%)	2,156 (3.97%)

Table 14 (Continued)

ICD-9 or ICD-10	Code	Description	Number of patients with code in total EHR database (N=181,931)	Number of patients with code in analysis cohort (N=54,259)
ICD-9	296.34	Major depressive affective disorder, recurrent episode, severe, specified as with psychotic behavior	3,128 (1.72%)	2,156 (3.97%)
	298.00	Depressive type psychosis	6 (0.0%)	2 (0.0%)
	300.40	Dysthymic disorder	8,642 (4.75%)	3,873 (7.14%)
	309.10	Prolonged depressive reaction	229 (0.13%)	94 (0.17%)
	296.00	Episodic mood disorders	182 (0.1%)	76 (0.14%)
	296.90	Other and unspecified episodic mood disorder	1,929 (1.06%)	1,136 (2.09%)

The eMERGE's 2/30/180 rule excluded 19.33% (N=35,173) of patients, leaving 146,758 patients. After filtering patients based on whether they were prescribed a drug from the inclusion medication list (Table 15), there were 54,259 patients in the final analysis cohort (Figure 8). We examined subsets of the analysis cohort to determine the proportion of patients who received behavioral therapy, underwent PGx testing, and had PHQ scores recorded (Figure 8).



	Number of patients
Therapy patients	12,539 (23.1%)
PHQ patients	3,726 (6.9%)
PGx patients	837 (1.5%)
Total cohort	54,259 (100%)

Figure 8 Modified CONSORT Diagram

5.3.2 Medication List Construction

There were 31 antidepressants included in the antidepressant therapeutic agent list (Table 15). Twenty medications were determined to be augmenting agents. Brand name drugs were converted into their generic drug equivalent. Drugs were converted to their antidepressant class for downstream analysis as well.

Table 15 Inclusion antidepressant list.

Type of agent	Class	Drug (Generic name)
Therapeutic agent (n=31)	SSRI	Sertraline
		Escitalopram
		Citalopram
		Fluoxetine
		Paroxetine
		Fluvoxamine
	SNRI	Desvenlafaxine
		Duloxetine
		Levomilnacipran
		Milnacipran
		Venlafaxine
	TCA	Amitriptyline
		Amoxapine
		Clomipramine
		Desipramine
		Doxepin
		Imipramine
		Nortriptyline
		Protriptyline
		Trimipramine
	MOA	Isocarboxazid
		Phenelzine
		Selegiline
		Tranylcypromine
	NDRI	Bupropion
	PHEN	Nefazodone
		Trazodone
	TET	Mirtazapine
		Maprotiline
	Miscellaneous	Vilazodone
		Vortioxetine
Augmenting agents and combinations	SNRI	Milnacipran
	Atypical agents	Aripiprazole
		Brexpiprazole
		Olanzapine
		Paliperidone
		Quetiapine
		Risperidone
	CNS stimulant	Armodafinil
		Atomoxetine
		Lisdexamfetamine

Table 15 (Continued)

Type of agent	Class	Drug (Generic name)
Augmenting agents and combinations	CNS stimulant	Methylphenidate
		Modafinil
	Anti-manic agent	Lithium
	Common combinations	Amitriptyline/chlordiazepoxide
		Amitriptyline/perphenazine
		Fluoxetine + olanzapine
	Miscellaneous	Esketamine
		L-methylfolate
		Niacin
		Thyroid desiccated

5.3.3 Electronic Phenotyping for Outcome Classification

There were 114,035 (77.70%) patients that were prescribed zero to one antidepressant, and therefore were considered to have “early depression” (Table 16). There were 32,723 (22.30%) patients that were considered to have “treatment-resistant depression” in having been prescribed two or more antidepressants.

Table 16 Number of unique antidepressants patients were prescribed.

Number of unique antidepressants prescribed	Number of patients (%)
0	92,499 (63.03%)
1	21,536 (14.67%)
2	17,400 (11.86%)
3	8,988 (6.12%)
4	3,932 (2.68%)
5	1,531 (1.04%)
6	575 (0.39%)
7	198 (0.13%)
8	74 (0.05%)
9	19 (0.01%)
>= 10	6 (~0.0%)
Total	146,758 (100%)

Of patients that had PHQ-9 scores documented within the EHR, most patients (N=460, 38.49%) were classified to have moderate depression upon their first encounter where a PHQ-9 score was documented (Table 17). Very few patients (N=27, 2.26%) were classified to have a severity level of none to minimal depression upon their first PHQ-9 score documentation encounter.

Among patients that had at least two PHQ-9 scores documented in the EHR (N=458 patients), the distribution of patients' last PHQ-9 score on record was similar to that of the distribution of initial PHQ-9 scores within the EHR. However, even fewer patients (N=4, 0.87%)

were classified to have none to minimal depression severity. When comparing patients' first PHQ-9 score to their last PHQ-9 score on record, the average score change was 0.76, which means that over the course of patients' treatment, scores increased slightly. The range of score changes was a decrease in 14 points and an increase in 14 points. The median score change was an increase by one point.

Table 17 Number of patients with varying levels of depression severity based on PHQ-9 score at their first and last PHQ-9 score encounter.

PHQ-9 score	Depression level severity	Number of patients with severity level at their **first** PHQ score collection encounter (N=1,195 patients)	Number of patients with severity level at their **last** PHQ score collection encounter (N=458 patients)
0-4	None to minimal	27 (2.26%)	4 (0.87%)
5-9	Mild	234 (19.58%)	98 (21.4%)
10-14	Moderate	460 (38.49%)	172 (37.55%)
15-19	Moderately severe	323 (27.03%)	139 (30.35%)
20-27	Severe	148 (12.38%)	45 (9.83%)

5.4 Discussion

Electronic phenotype construction is paramount to transportability and reproducibility. Studies using EHR data require electronic phenotypes in order to conduct reproducible translational research, comparative effectiveness studies, CDS work, and public health studies [119]. In this study, the electronic phenotype for MDD from PheKB was implemented, along with the 2/30/180 rule to minimize the potential for MDD ICD-9/-10 codes being present in the EHR as a result of administrative errors. A majority of patients in the cohort had ICD-10 codes for MDD within their EHR, over MDD ICD-9 codes.

Electronic phenotypes for outcome classifications of treatment response and symptom remission were also implemented for patients that had PHQ scores documented within their EHR. The majority of patients were classified to have moderate depression at the onset of their first PHQ score documentation encounter. The distribution of depression severity was similar at patients' last documented PHQ score encounter. Future directions include mapping out the trajectories at a more granular level between the first and last documented PHQ scores for patients.

6.0 Characterizing Antidepressant Prescribing Sequences

6.1 Introduction

The Observational Health Data Science and Informatics (OHDSI) collaboration has looked at the sequence of drugs prescribed for patients in order to look at treatment care pathways for diabetes, hypertension, and depression [129]. Namely, 11% of depression patients had a unique treatment pathway.

The sequence of disease progression pathways is also of interest in examining clinical care data. Kwon et al. (2020) used Hidden Markov Models (HMMs) in order to map disease progression pathways [130]. HMMs are useful due to their ability to ascertain latent states and transition states between ascertained states using time variable, incomplete, missing, and irregular multivariate data. The longitudinal nature of treatment pathways makes HMMs an ideal method for modeling the longitudinal data. In addition, HMMs are able to incorporate the uncertainties inherent to clinical data.

Sukkar et al. (2012) used HMMs to model disease progression as opposed to clinical stages of disease, and were able to elicit greater granularity in disease stages from their HMM model as opposed to conventional clinical stages of disease [131]. Sampathkumar et al. (2014) used a HMM to extract adverse drug reactions from text [132]. Chen et al. (2019) implemented an HMM-based method to improve prediction of a disease state that signals disease progression towards AIDS [133].

Liu et al. (2015) developed a continuous-time HMM to model disease progression with temporal data consisting of irregularly sampled time points [134]. Sun et al. (2019) implemented

a continuous-time HMM to model observational data to represent Huntington's disease progression in patients [135]. Kwon et al. (2020) used HMMs with interactive visualizations in order to display disease progression for type 1 diabetes, Huntington's disease, Parkinson's disease, and chronic obstructive pulmonary disease [130].

For this study, due to the fact that the majority of patients did not have a large number of depression severity score data to determine depression severity phenotypes for patients over the course of their treatment, a Markov chain model was implemented to examine the transition probabilities between antidepressants that patients were prescribed. A Markov chain is a probabilistic model that operates under the dependence that the future does not depend on the past [136]. Given a sequence of random variables in the state space (or antidepressant state space), a Markov chain model is conveyed through the conditional probabilities.

6.2 Methods

6.2.1 Characterize patient experience with antidepressants

For all MDD patients that were prescribed an antidepressant, the time between the first MDD ICD-9/-10 code and the first antidepressant prescribed was calculated and plotted. The time intervals between first MDD ICD-9/-10 code and first antidepressant prescribed were also analyzed according to gender and race, and a one-way ANOVA and a Tukey's Honest Significant Difference (HSD) test was used to compare group means between races.

6.2.2 Display Sequence of Antidepressants Prescribed

Patients' antidepressant treatment sequences, in terms of medication prescription events, were extracted from the EHR database. Antidepressants were then further annotated according to the antidepressant class and the drug's generic name equivalent. Unique antidepressant treatment sequences, according to individual antidepressant and antidepressant class, were then tallied according to how many patients followed the same treatment path. The number of unique antidepressants prescribed was recorded. The number of antidepressants prescribed over each patient's EHR history was normalized according to age and time since initial diagnosis. Patients with a diagnosis code for depression that were never prescribed an antidepressant will also be tallied and excluded from model construction.

6.2.3 Sequence of Antidepressants

The inclusion criteria included a list of antidepressant medications (Table 15) and was used to search through the EHR database for patients' antidepressant prescription history with antidepressants. Antidepressant prescription events were then put in consecutive order for each patient according to the prescription event date. Brand name antidepressants were converted to their generic name to allow for parallel comparison between drugs. Antidepressants were also converted to their class name to examine antidepressant prescribing sequences at the class level.

When examining the sequence of antidepressants that patients were prescribed, repeated prescription events of the initial antidepressant were not recorded, only the additional antidepressant was captured when looking at the antidepressant sequence level. However, when looking at antidepressant prescription events for whether there was a medication switch,

continuation, dose change, or addition of augmentative therapy, antidepressants prescribed were unaltered and patients' subsequent antidepressant prescription events were captured.

Patients were prescribed their second antidepressant at variable time frames. Second antidepressant prescription events were characterized in relation to the initial antidepressant prescription event. Second antidepressant prescription events were either a drug switch or continuation of the drug. Continuations were further classified according to whether there was a dose change or continuation, and whether an augmenting agent was prescribed in addition.

6.2.4 Markov Model of transition probabilities between antidepressants

With all of the antidepressant treatment sequences for each patient, Markov chain models were constructed for generic antidepressant medication orders and antidepressant classes using the package 'markovchain' in R. Markov Models were deemed appropriate in order to examine the transition probabilities between being prescribed certain antidepressants due to the fact that prescribing for depression is currently a trial-and-error process. A figure conveying transition probabilities was constructed. The transition probability matrix (P) was a square matrix that took the following form:

$$P = \begin{bmatrix} \rho_{11} & \rho_{..} & \rho_{1s} \\ \rho_{..} & \cdots & \cdots \\ \rho_{s1} & \cdots & \rho_{ss} \end{bmatrix}$$

where s was the total number of antidepressants, and p was the probability of a patient going from being prescribed one antidepressant (rows) to another antidepressant (columns). Therefore, all rows of P summed to one.

6.3 Results

6.3.1 Characterization of patients' experience with antidepressants

The time intervals between the first MDD ICD-9/-10 code within a patient's EHR and the first antidepressant prescribed ranged from -1,058 days to 2,783 days, with an average time interval of 438 days (Figure 9). When patients' time intervals were compared according to gender, males had a longer average time interval at 553 days compared to females at 397 days (Figure 10). A two-sample z-test revealed that the mean difference in time intervals between males and females. was statistically significant ($P < 0.0001$). In addition, patients' median time intervals according to race were compared, which revealed that the median time interval was lowest for individuals that identified as White at 553 days (1.51 years), and the median time interval was greatest for patients that identified as Asian Pacific Islander at 1,477 days (4.04 years) (Figure 11). A one-way ANOVA determined that there was a statistically significant difference between the time intervals for race ($P < 0.0001$), and a Tukey HSD test comparing means revealed that there was a statistically significant difference between Black and Asian Pacific Islander patients ($P < 0.0001$), White and Asian Pacific Islander patients ($P < 0.0001$), Unknown race and Black patients ($P < 0.0001$), White and Black patients ($P < 0.0001$), and Unknown race and White patients ($P < 0.0001$).

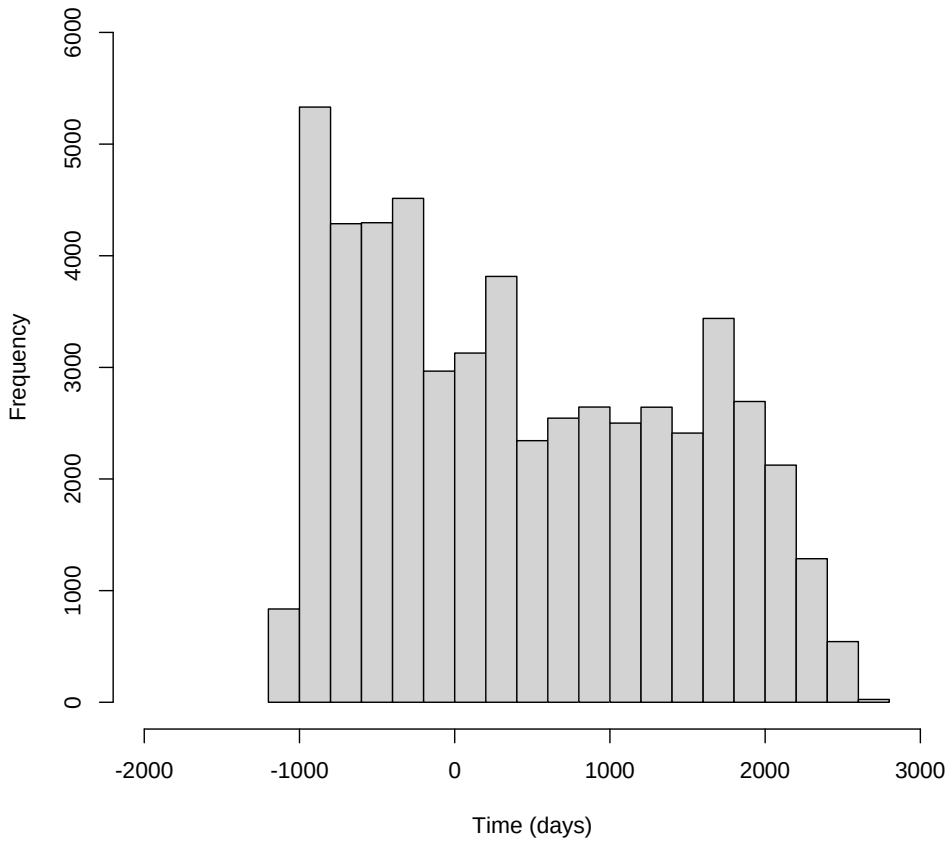


Figure 9 Histogram of time between patients' first MDD ICD-9/-10 code and first antidepressant prescribed.

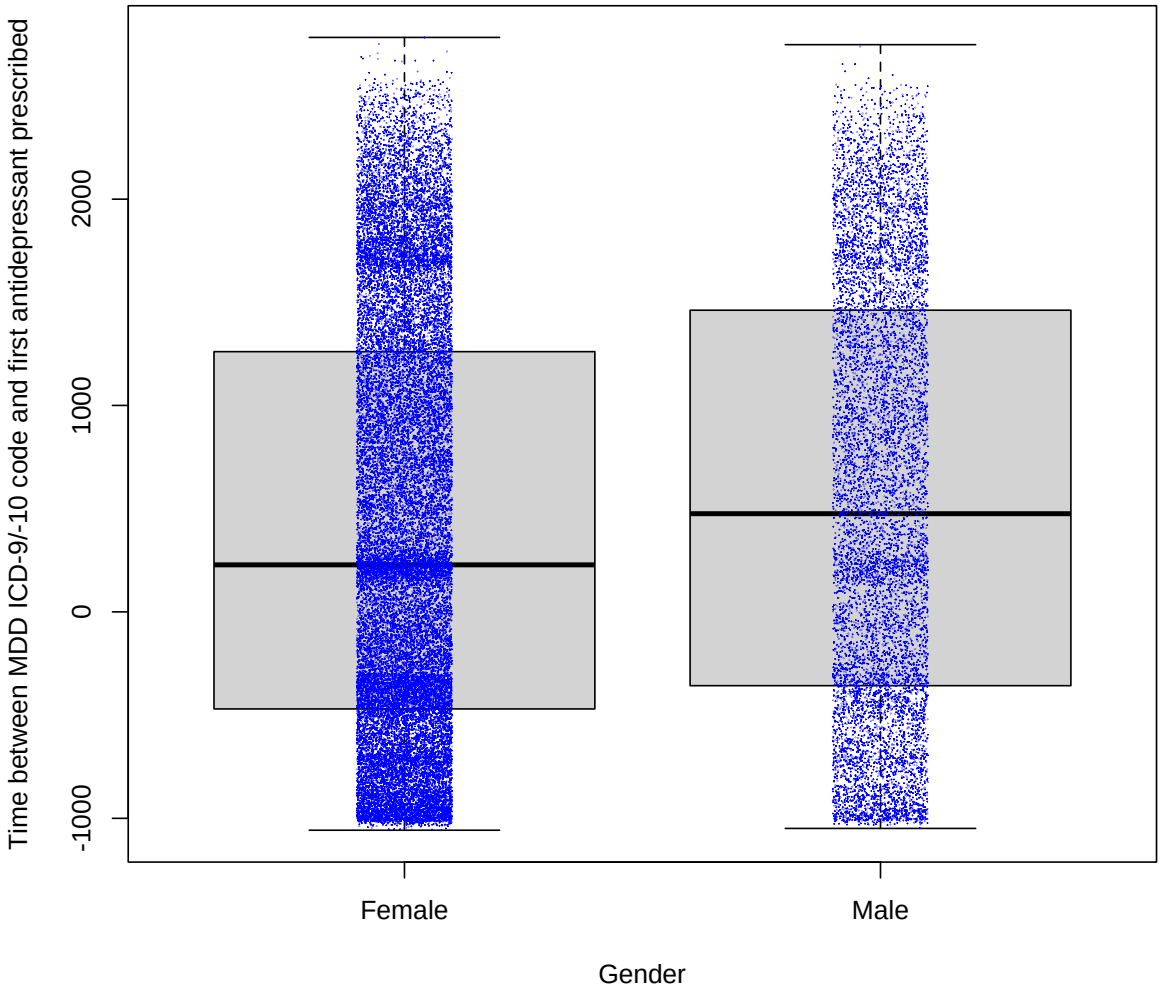


Figure 10 Boxplots of time between ICD-9/-10 codes and first antidepressant prescribed stratified according to gender.

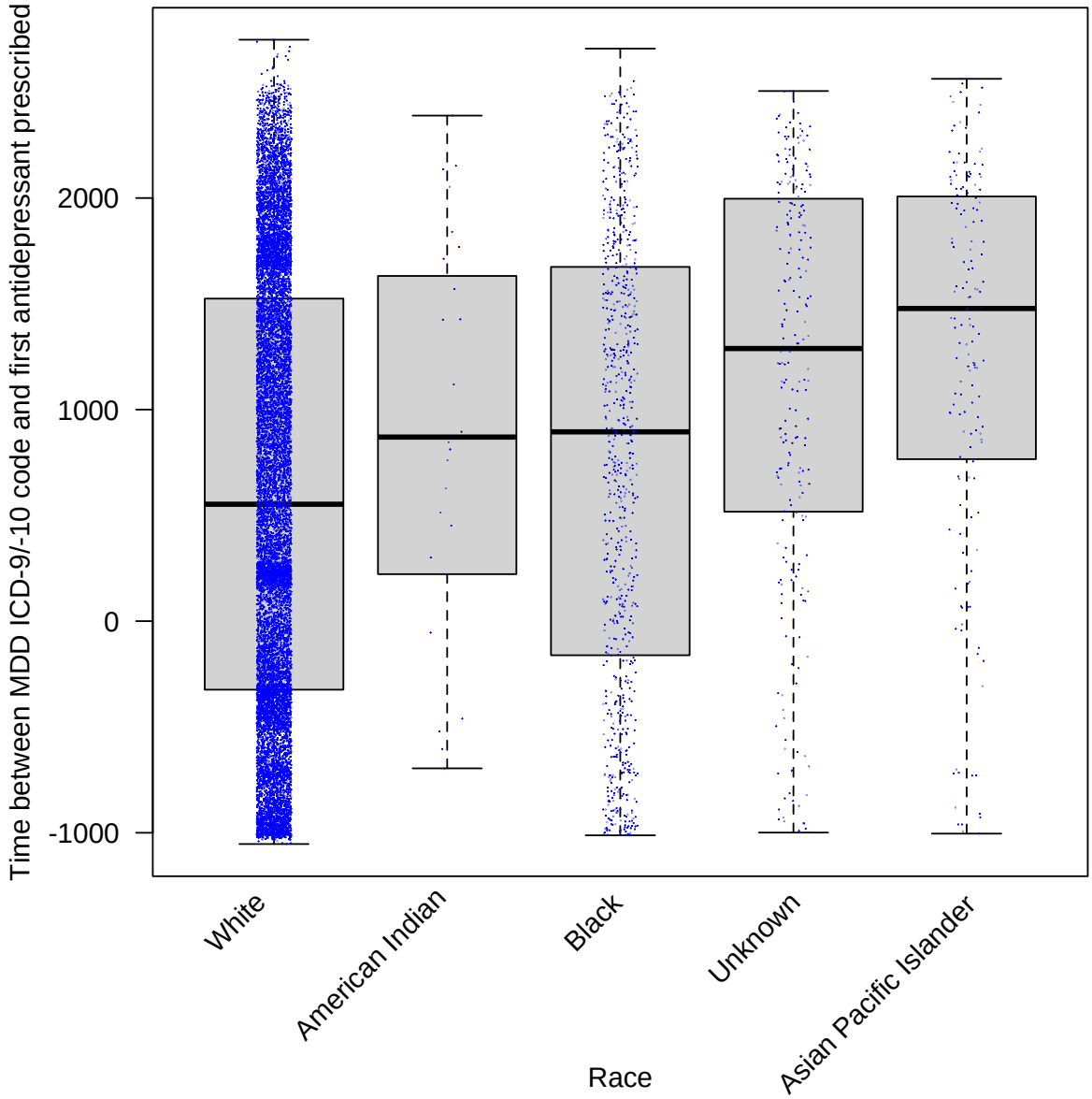
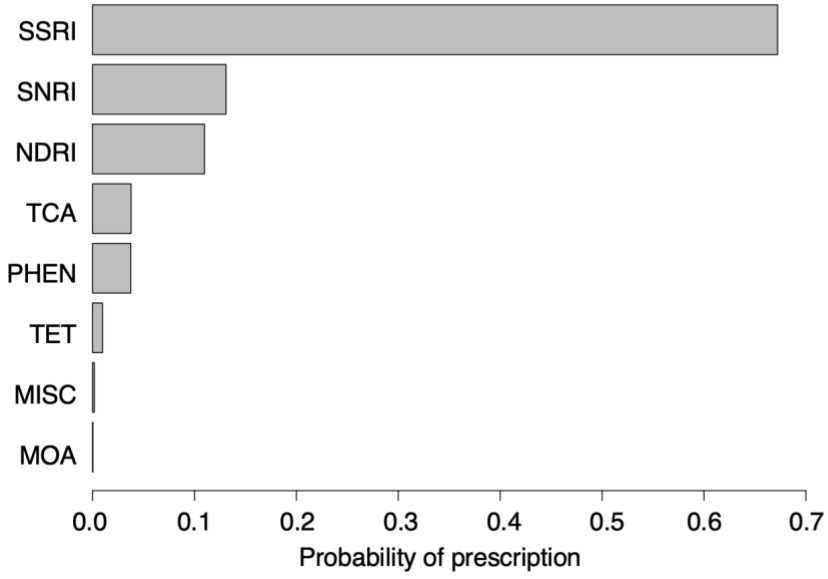


Figure 11 Boxplots of time between ICD-9/-10 codes and first antidepressant prescribed stratified according to race.

6.3.2 Sequence of Antidepressants Results

First-line prescribed antidepressants were mostly SSRIs (67.22.0%), followed by SNRIs (13.10%) and NDRIs (10.98%) (Figure 12). The most frequently prescribed initial antidepressants

were SSRIs sertraline (20.32%), citalopram (17.32%), and escitalopram (12.55%) (Figure 13). Of patients that were prescribed at least two antidepressants (N=34,894, 64.31%), the most frequent antidepressant prescribing sequence was citalopram then bupropion (3.48%), followed by sertraline then bupropion (3.38%), and citalopram then sertraline (3.10%) (Figure 14). There were 425 unique two-drug prescribing sequences of generic drugs. For patients that were prescribed at least three antidepressants (N= 27,156, 50.05%), there were 6,274 unique prescribing sequences. For these patients that were prescribed at least three drug antidepressants, the most prevalent prescribing sequence was escitalopram, duloxetine, then escitalopram (1.52%), followed by bupropion, citalopram, then escitalopram (1.28%), and bupropion, escitalopram, then duloxetine (1.00%) (Figure 15).



Class Label	Class
SSRI	Selective serotonin reuptake inhibitors
SNRI	Serotonin and norepinephrine reuptake inhibitors
NDRI	Norepinephrine-dopamine reuptake inhibitors
TCA	Tricyclic antidepressant
PHEN	Phenylpiperazine
TET	Tetracyclic antidepressant
MISC	Miscellaneous
MOA	Monoamine oxidase inhibitors

Figure 12 Initial antidepressant class prescription probabilitiy for UPMC depression patients.

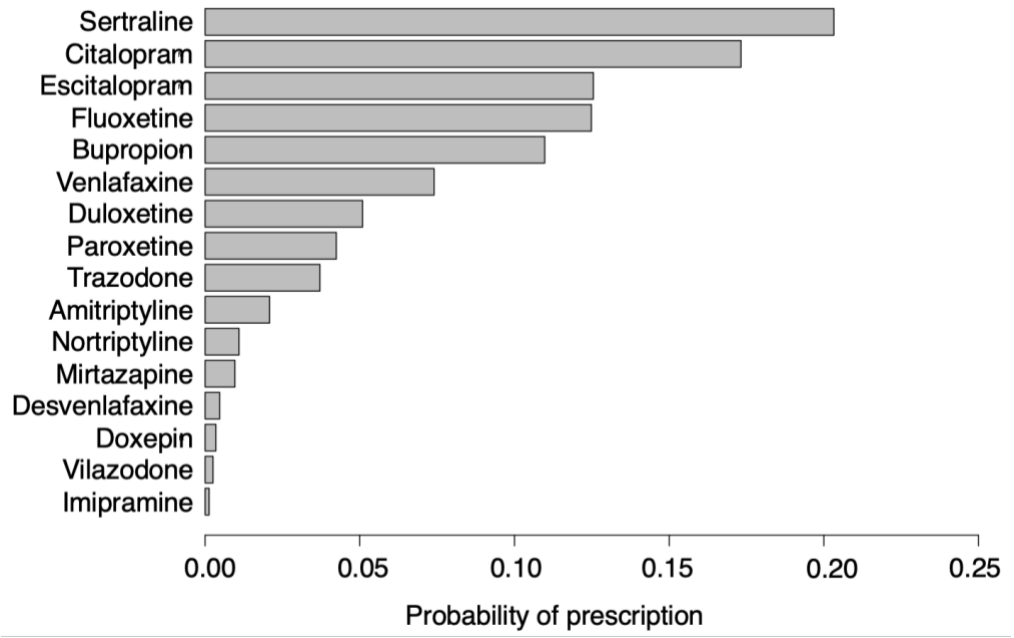


Figure 13 Top 15 initial antidepressants prescription probabilities for UPMC depression patients.

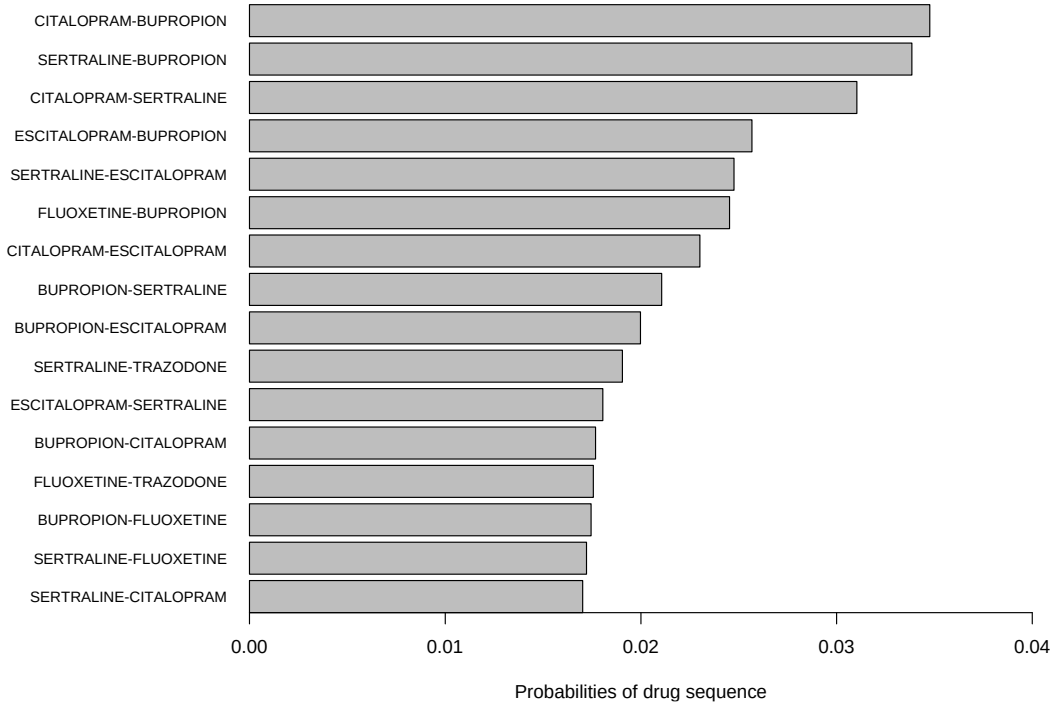


Figure 14 Top 15 two-drug antidepressant sequences for UPMC depression patients.

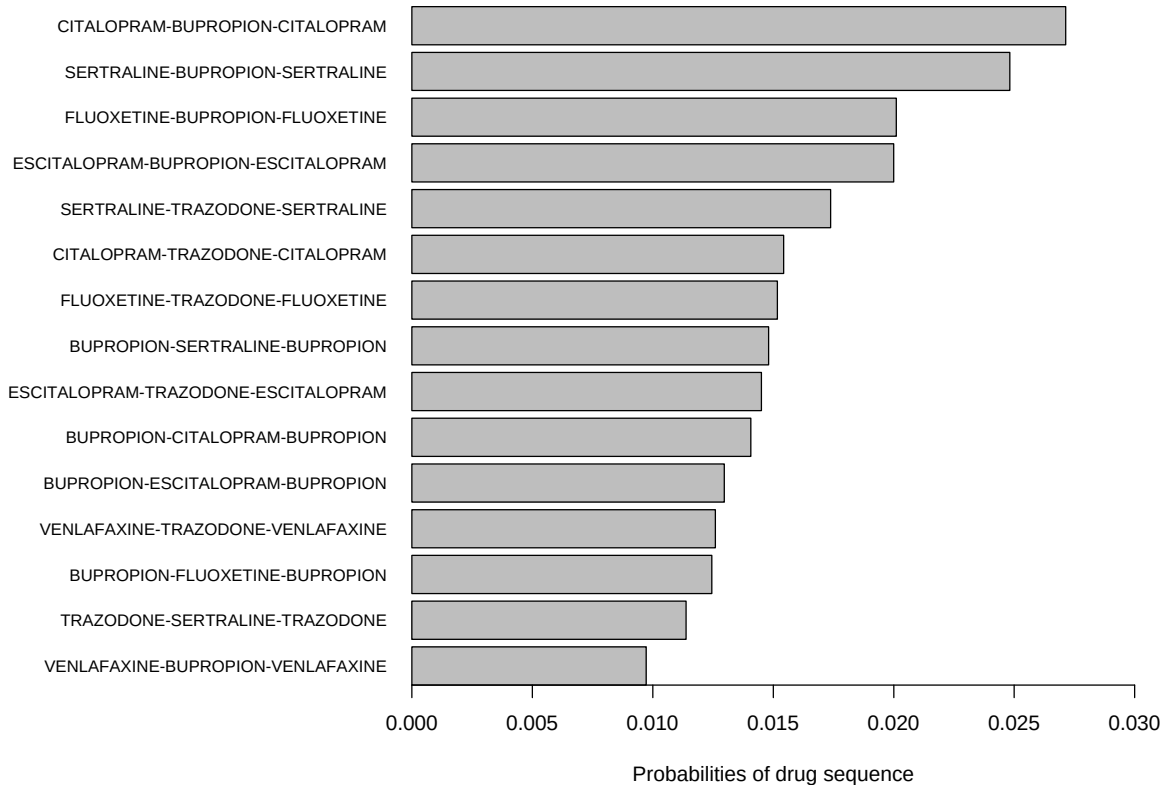


Figure 15 Top 50 three-drug antidepressant sequences for UPMC depression cohort patients.

When examining polypharmacy, there were 78,519 instances of antidepressant combinations prescribed for patients in the analysis cohort across their history of antidepressant prescription events. The most prevalent antidepressant combination for an antidepressant prescription event was bupropion and sertraline prescribed together (n=3,051, 3.89%), followed by bupropion and escitalopram (n=3,032, 3.86%), and bupropion and fluoxetine (n=2,981, 3.80%) (Table 18).

Table 18 Top 15 most prevalent polypharmacy prescriptions.

Drug combination	Number of instances (%)
Bupropion + sertraline	3,051 (3.89%)
Bupropion + escitalopram	3,032 (3.86%)
Bupropion + fluoxetine	2,981 (3.80%)
Bupropion + citalopram	2,917 (3.72%)
Trazodone + sertraline	2,417 (3.08%)
Bupropion + venlafaxine	1,957 (2.49%)
Bupropion + trazodone	1,770 (2.25%)
Trazodone + citalopram	1,752 (2.23%)
Venlafaxine + trazodone	1,534 (1.95%)
Fluoxetine + trazodone	1,382 (1.76%)
Bupropion + duloxetine	1,444 (1.84%)
Escitalopram + trazodone	1,295 (1.65%)
Duloxetine + trazodone	930 (1.18%)
Bupropion + paroxetine	584 (0.74%)
Mirtazapine + sertraline	569 (0.72%)

Most (52.29%, N=28,31) second antidepressants were prescribed within six weeks after the initial antidepressant prescription event (Table 19). Across all time frames for the second antidepressant prescribed, most (65.63%, N=35,687) patients continued their initial antidepressant drug and dose at the second prescription event. Very few (0.65%, N=351) patients were prescribed an additional augmenting agent along with the initial antidepressant prescribed at the second

antidepressant prescription event. When the second antidepressant prescription event involved a dose change, the second prescription event was more often a dose increase (83.41%) over a decrease (16.59%). A prescription for a dose increase or additional augmenting agent might have signified incomplete activity, while a dose decrease might have signified toxicity. Intolerance might be gleaned from patients that switched drugs upon the second antidepressant prescribed or decreased dose, which occurred less frequently in the first two weeks (10.08%, N=2,087 and 0.44%, N=91, respectively) and as might be expected, increased in frequency as the time to the second antidepressant prescription event increased. In total, 2,178 (10.5%) patients that had a second prescription encounter in the first two weeks could have been intolerant of the drug based on the medication switch or decrease in dose.

Table 19 Characterization of the subsequent prescription event after the initial antidepressant prescription event for patients.

Time to second antidepressant prescribed	Switch drug	Continue drug				Total
		Continue dose	Change dose		Add augmenting agent	
			Increase dose	Decrease dose		
[0 weeks, 2 weeks)	8,989 (43.43%)	11,072 (53.49%)	466 (2.25%)	95 (0.46%)	76 (0.37%)	20,698 (38.07%)
[2 weeks, 4 weeks)	697 (20.42%)	1,640 (48.04%)	864 (25.31%)	71 (2.08%)	142 (4.16%)	3,414 (6.28%)
[4 weeks, 6 weeks)	769 (17.81%)	2,239 (51.84%)	1,055 (24.43%)	91 (2.11%)	165 (3.82%)	4,319 (7.94%)
[6 weeks, 8 weeks)	542 (22.71%)	1,237 (51.82%)	478 (20.03%)	50 (2.09%)	80 (3.35%)	2,387 (4.39%)
[8 weeks, 3 months)	710 (21.83%)	1,932 (59.41%)	443 (13.62%)	82 (2.52%)	85 (2.61%)	3,252 (5.98%)
[3 months, 6 months)	1,345 (21.56%)	3,996 (64.06%)	611 (9.79%)	152 (2.44%)	134 (2.15%)	6,238 (11.47%)
[6 months, 12 months)	1,517 (26.86%)	3,415 (60.46%)	413 (7.31%)	170 (3.01%)	133 (2.35%)	5,648 (10.39%)
[12 months, ∞)	4,162 (49.44%)	3,096 (36.78%)	610 (7.25%)	374 (4.44%)	176 (2.09%)	8,418 (15.48%)
Total	18,731 (34.45%)	28,627 (52.65%)	4,940 (9.08%)	1,085 (2.00%)	991 (1.82%)	54,374 (100%)

6.3.3 Markov Model Results

Markov chain models revealed that transition probabilities from any class of antidepressant to SSRIs were highest ($\bar{x}$ = 0.83, median = 0.83). Transition probabilities from any class of antidepressant to SNRIs ($\bar{x}$ = 0.11, median = 0.11) and to NDRIs were low ($\bar{x}$ = 0.03, median =

0.03) (Figure 16, Figure 17). When examining transition probabilities between individual antidepressants, antidepressant class transition probability findings were replicated in that transition probabilities were highest for transitioning to SSRIs. The highest transition probabilities were switching from another antidepressant to sertraline ($\bar{x}$ = 0.17, median = 0.17), fluoxetine ($\bar{x}$ = 0.14, median = 0.14), escitalopram ($\bar{x}$ = 0.16, median = 0.14), and citalopram ($\bar{x}$ = 0.14, median = 0.14) (Figure 18). These SSRIs were also the antidepressants that had the highest probability of being prescribed as initial therapies. Transition probabilities between individual SSRI antidepressants were not higher than transitioning from an SSRI to another antidepressant (Figure 18). A few transition probabilities were artificially high due to low N, namely transitioning from isocarboxazid to paroxetine (N = 4, 66%) and amoxapine to sertraline (N = 3, 60%). Overall, transition probabilities were highest from any class to SSRIs, and from any individual antidepressant to the SSRIs sertraline, fluoxetine, escitalopram, and citalopram. Therefore, SSRIs were not only found to be used as initial therapy options, but are also revisited as therapy options.

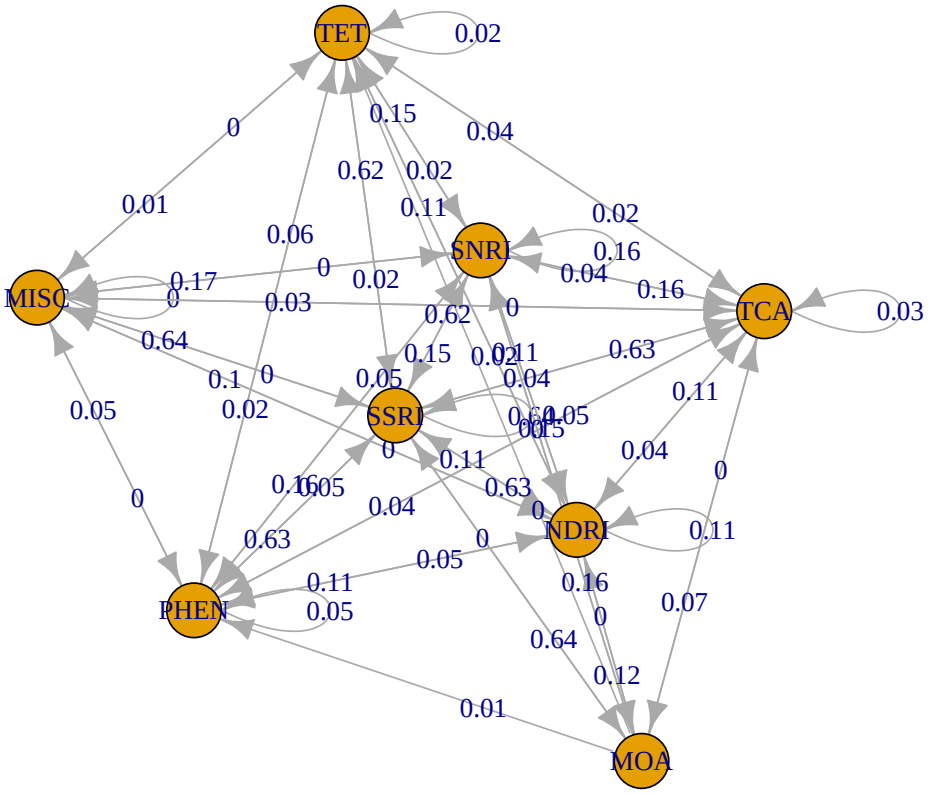


Figure 16 Markov Model of antidepressant classes prescribed for UPMC depression patients.

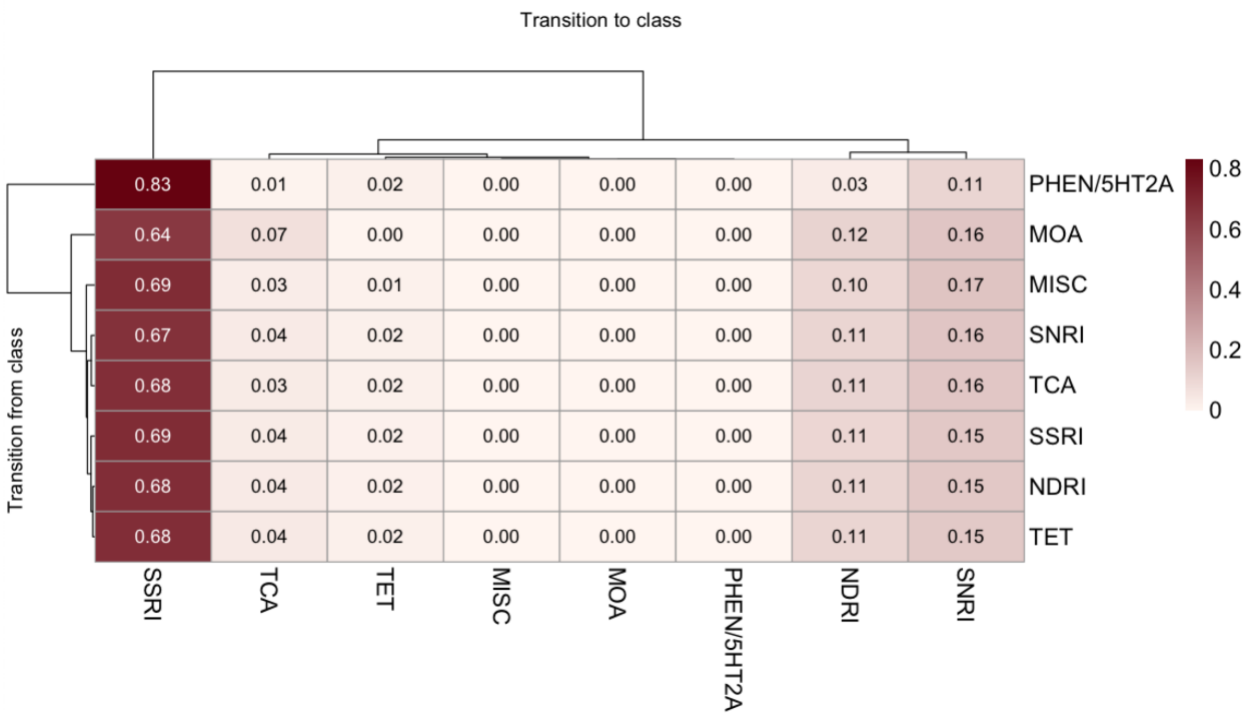


Figure 17 Transition probabilities between antidepressants classes

phenylpiperazine / 5-HT2 receptor antagonists (PHEN/5HT2A), monoamine oxidase inhibitors (MOA), miscellaneous (MISC), serotonin and norepinephrine reuptake inhibitors (SNRI), tricyclic antidepressants (TCA), selective serotonin reuptake inhibitors (SSRI), norepinephrine-dopamine reuptake inhibitors (NDRI), and tetracyclic antidepressants (TET) prescribed for UPMC depression patients.

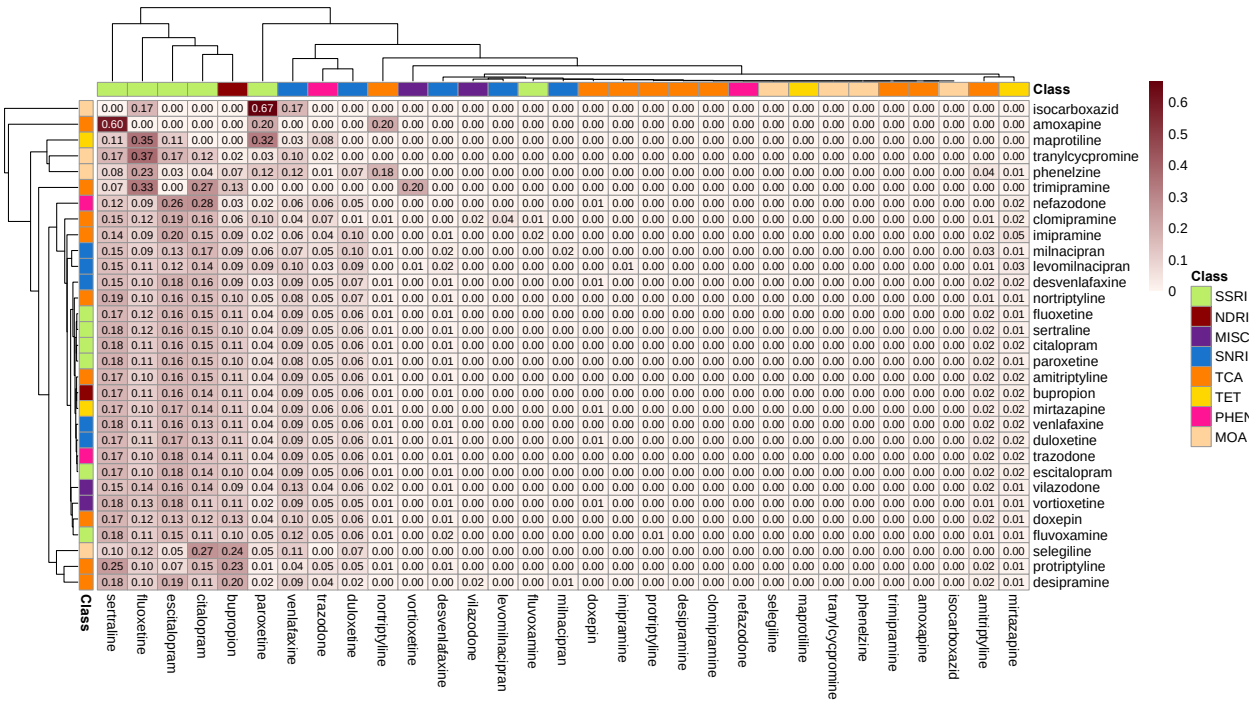


Figure 18 Transition probabilities between individual antidepressant drugs prescribed, annotated by antidperessant class

selective serotonin reuptake inhibitors (SSRI), norepinephrine-dopamine reuptake inhibitors (NDRI), miscellaneous (MISC), serotonin and norepinephrine reuptake inhibitors (SNRI), tricyclic antidepressants (TCA), tetracyclic antidepressants (TET), phenylpiperazine / 5-HT2 receptor antagonists (PHEN/5HT2A), and monoamine oxidase inhibitors (MOA), where the y-axis represents transitioning from an antidepressant, and the x-axis represents transitioning to an antidepressant.

Figure 18 did not convey that patients were more likely to be prescribed a different SSRI subsequently after another SSRI. Transition from one SSRI to another SSRI was not significantly higher than switching from another antidepressant class to an SSRI. However, when a clinician switches the prescription order within SSRIs, sertraline (the most frequently prescribed initial antidepressant) had the highest transition probability ($\bar{x}$ = 0.28, median = 0.28), and therefore was the most likely SSRI to be prescribed after prescribing another SSRI (Figure 19, Figure 20).

Escitalopram had the next greatest mean transition probability ($\bar{x}$ = 0.25, median = 0.25). The lowest transition probability for intra-SSRI transitions was for fluvoxamine ($\bar{x}$ = 1.58 x 10^{-3}, median = 1.86 x 10^{-3}), followed by paroxetine ($\bar{x}$ = 6.52 x 10^{-2}, median = 6.04 x 10^{-2}).

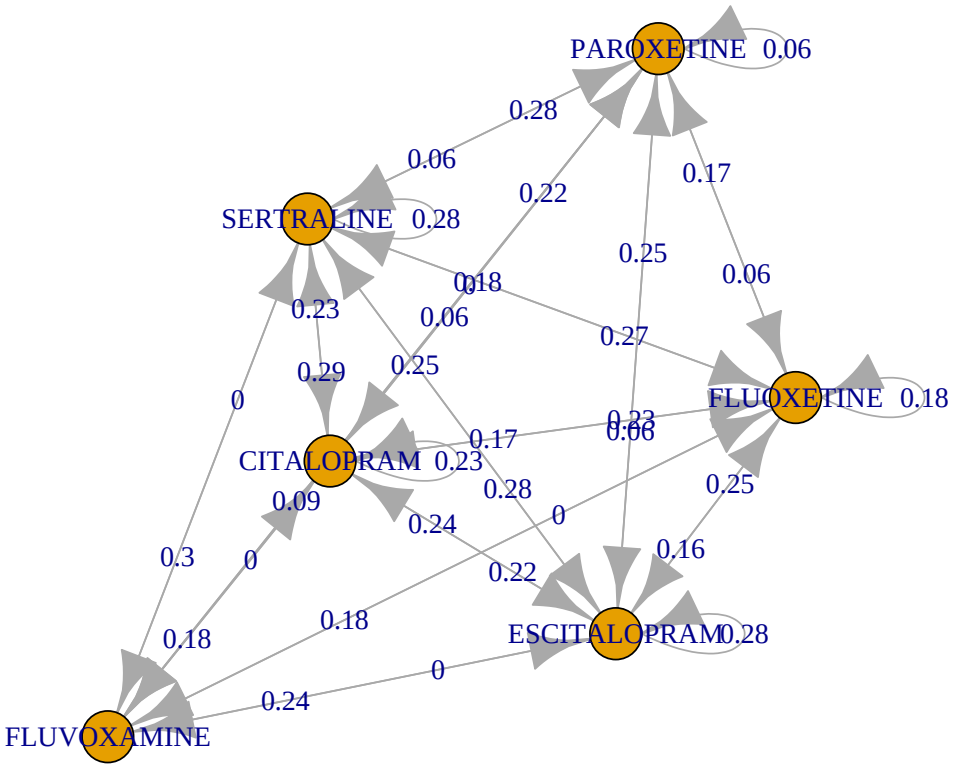


Figure 19 Markov Model of intra-SSRI transition probabilities for UPMC depression patients.

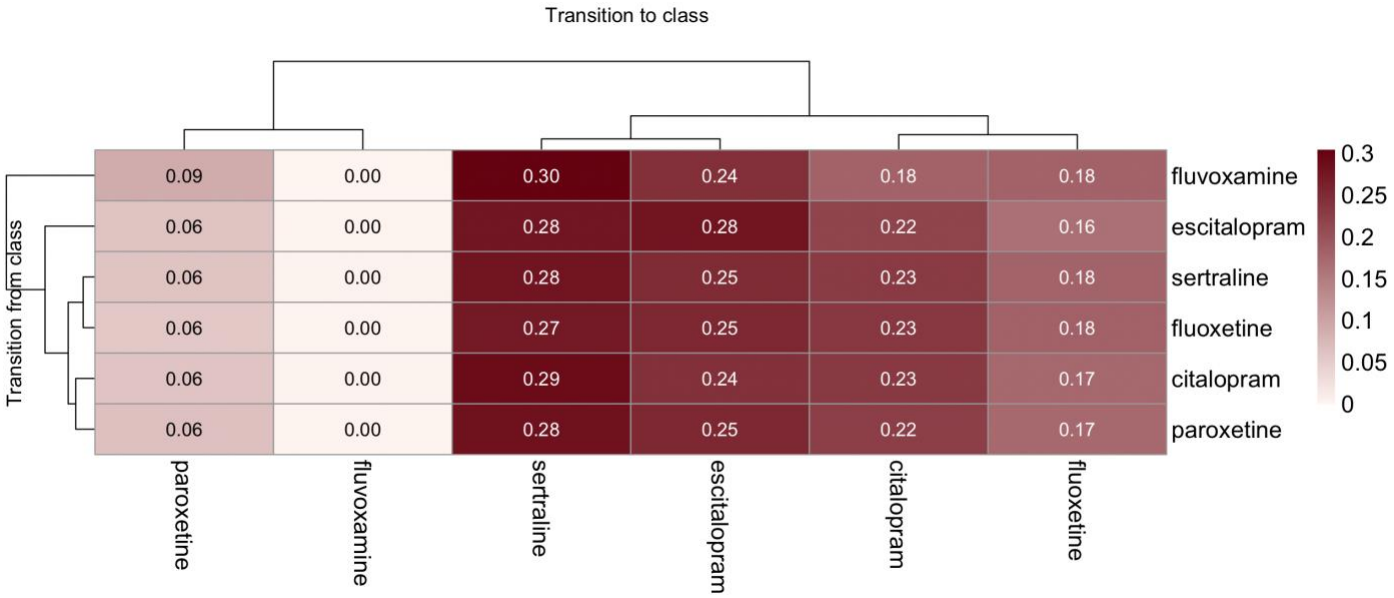


Figure 20 Transition probabilities for intra-SSRI prescribing for UPMC depression patients.

6.4 Discussion

The time intervals between patients' first MDD ICD-9/-10 code and patients' first prescribed antidepressant disparities according to gender and race were significant and therefore are potential points for clinical care improvement. There are many factors that could impact the time between patients' first MDD ICD-9/-10 code and when the patient might be prescribed an antidepressant, however this is an area that could be investigated further. Especially as it relates to the role that culture and gender play in how depression is treated. There were no studies identified concerning the time between initial MDD diagnosis and first antidepressant prescription, and how it relates to patients' demographic factors.

Milea et al. (2010) [137] used a U.S. claims database to convey initial antidepressants prescribed for patients, and found very similar results, with most patients being prescribed

sertraline as initial therapy, followed by escitalopram, bupropion, paroxetine, and fluoxetine. In addition, Olekhnovitch et al. (2020) [138] examined efficacy and tolerability of antidepressants through looking at medications fills data to inform first-line medications. This study found that the ratio of medication continuations over changes was highest for escitalopram, followed by sertraline, venlafaxine, citalopram, fluoxetine, and paroxetine. Future directions for this work could also subset medication switches and continuations based on the initial antidepressant prescribed for each patient.

Sawada et al. (2009) [139] found in a chart review of Japanese patients prescribed antidepressants, only 44.3% of patients continued antidepressant treatment for 6 months, and found that sertraline was continued at the highest rate at 6 months over other antidepressants. In the UPMC EHR database, 72.4% of patients continue their antidepressant treatment within the clinical time frame of 6 months, which is an interesting finding concerning patients' compliance and adherence.

There were 10.5% of patients that could have been intolerant of their initial antidepressant prescribed based on a medication switch or dose decrease within the first two weeks of their original prescription. These patients represent a potential utility for PGx data in order to uncover poor or ultra-rapid metabolizers before the initial antidepressant prescription. This could allow for a shorter time to identify a successful antidepressant for the patient.

There were no studies identified that constructed a Markov Chain model for antidepressants prescribed, and therefore these results were unable to be compared to the literature base. However, this may connote a novel scholarly contribution to enhance the understanding of antidepressant prescribing patterns, especially in terms of the high transition probabilities to prescribing SSRIs outside of just initial first-line therapy prescriptions. When a patient does not

respond or partially responds to the initial antidepressant, clinicians have the option of increasing the dose, augmenting with another agent, or switching to another antidepressant. Therefore, high transition probabilities to SSRIs is particularly interesting. Transition probabilities were not higher within SSRIs, so patients were not more likely to switch to yet another SSRI after being prescribed a different SSRI. Future directions of this work could look at Markov Chain models that include augmentation and dose changes in addition to drug switches and continuations.

7.0 Modeling Treatment Response and Symptom Remission

7.1 Methods

7.1.1 Feature characterization and feature selection

Features from the demographics, diagnoses, encounters, medication fills, medication orders, vitals, and questionnaire data were used to create the training and test set for the machine learning models. The individual antidepressants that patients were prescribed within the first twelve weeks were not included in the model, however the majority of patients (62.45%) were prescribed one SSRI as first-line therapy. Mutual information, a measurement of the relationship between two random variables, was calculated between each feature and the outcome variable. Mutual information is calculated between two random variables, X and Y as follows [140]:

$$I\ (X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

where P(x) and P(y) are the marginal distributions of X and Y.

Therefore, mutual information conveys how close the joint distribution of two variables is to the independent joint distribution. If X and Y are independent, then P(x,y) = P(x)P(y) and the mutual information score would be zero. Mutual information is also a useful metric for this dataset because it can be calculated on multiple univariate data types, for example, continuous, binary, and categorical data.

A priori models were created to predict both treatment response and symptom remission based on sociodemographic features implemented in MDD published literature, namely age,

gender, race, ethnicity, and BMI. A priori models for treatment response and symptom remission were then compared to prediction models using K-best features, where K=5, 10, 25, and 50, based on mutual information scores. Pearson correlation coefficients were calculated between each of the top features in each feature set. A threshold of 0.85 was established so as to not allow two features that were too highly correlated. When features had a Pearson correlation coefficient greater than 0.85, the feature with the greater mutual information score was kept in the feature space, and the feature with the lower mutual information score was dropped. In order to maintain a consistent number of features in the space, the feature with the next highest mutual information score was then added to the model feature space.

7.1.2 Machine Learning Models to Predict Successful Antidepressants

An a priori models using features from published literature were run using logistic regression. Once K-best feature selection based on mutual information scores was run, separate models for the binary predictions of symptom remission and treatment response for patients between zero and twelve weeks were constructed using logistic regression, random forest, and stacked ensemble models to determine the model with the greatest performance. Performance was assessed through common performance metrics calculated from confusion matrices including, sensitivity, specificity, positive predictive value, negative predictive value, and F-score.

7.1.2.1 Logistic regression

$$y_i = \log\left(p\frac{p}{1-p}\right)$$
$$= \beta_0 + \beta(Age) + \beta(Gender) + \beta(Race) + \beta(Ethnicity)$$
$$+ \beta(Behavioral\ Therapy) + \cdots + \beta(Clinical\ Feature_N)$$

Logistic regression models were formed using the top 5, 10, 15, and 25 features for predicting treatment response, and the top features for predicting symptom remission. The logistic regression model computes the log odds of predicting whether a patient experienced treatment response or symptom remission. Each coefficient in the model corresponds to the relative increase in log odds of a particular predicted outcome, where one unit increase in a feature X_i for a particular patient leads to a β increase in the log odds of a patient experiencing treatment response ($y_i = 1$, for a patient having treatment response), for example. The maximum likelihood estimation computes the coefficients through an optimization of the likelihood function, which consists of the probability of observing the patient data given the logistic regression model coefficients. The probability of the likelihood function is maximized when the coefficients compute predicted outcomes that are closest to the actual outcomes for patients. Odds ratio, standard error, and P-value for each feature were reported.

7.1.2.2 Random forest

Random forests are an ensemble of classifiers $h_1(x), h_2(x), \ldots, h_K(x)$, where each classifier is a decision tree. A decision tree uncovers features with the greatest discrimination in classifying samples according to samples' labels. A decision tree splits samples at each node according to the feature that maximizes the split in samples. Random forests are an ensemble of

decision trees that use training data $\{\{(x_i, y_i)\}_{i=1}^{N}$ drawn randomly with replacement from X and Y to create an ensemble of trees where the larger the margin in feature classification is selected for splitting at each node. Random forests construct trees on subsamples of the data and in addition, randomize the feature set available at each node to select the feature with the best split. Then the ensemble of decision trees in the random forest is tested on the hold-out set to determine model performance [141].

Random forest models were constructed on the datasets containing the top 5, 10, 25, and 50 features subset randomly with replacement to create an ensemble of 500 trees. Hyperparameter tuning was conducted to optimize the number of trees in the forest, the maximum number of terminal nodes in the forest, and the number of candidates to draw from to run the algorithm. Five-fold cross-validation was conducted as well, to improve generalizability across models.

7.1.2.3 Auto ML

The H2O Automatic Machine Learning (AutoML) package [142] was used to train machine learning algorithms contained within the package like generalized linear modeling (including logistic regression), decision trees (including gradient boosting machine, XGBoost, random forest), and deep learning frameworks. Generalized linear models iteratively weight coefficients in calculating maximum likelihood estimates and perform transformation to coerce a linear relationship [143].

Gradient boosting machine trains decision trees successively one at a time and weights trees based on a loss function that measures the difference between patients' predicted outcomes and patients' actual outcomes [144]. Trees are added to the ensemble that ultimately minimize the loss function. A benefit of gradient boosting machine is that the loss function can be specified, and it is typically high performing in its combination of many decision trees. However, drawbacks to

gradient boosting machine are that resulting models can be very complex and difficult to interpret, it can be difficult to perform hyperparameter tuning, and individual trees can lead to overfitting and therefore a lack of generalization to unseen data. Overfitting can be overcome through hyperparameter tuning like adjusting the number of trees, tree depth, number of split nodes, and the number of samples per split. A type of gradient boosting machine is XGBoost, or eXtreme Gradient Boosting, and was designed to optimize performance and speed [145]. XGBoost also implements the mechanism called boosting where new decision trees are trained in succession to minimize the loss function between predictions and actual outcomes. A strength of XGBoost is its ability to handle missing data, in addition to implementing parallel and distributed computing to improve speed. However, XGBoost has the same limitations as a gradient boosting machine in that it can overfit data, be challenging to tune, and result in highly complex trees that are difficult to interpret.

Another ensemble tree method used by the H2O package is the distributed random forest that trains multiple decision trees on random subsets of patients and features in the dataset and computes the average of decision trees [141]. Random forest does not implement boosting to adjust the weights of trees, instead subsets of patients and features are used to train multiple trees in parallel that are combined in an ensemble in a process called bagging. Random forests are typically high performing, are robust to outliers and missing data, however they are also difficult to interpret, can overfit, and are computationally expensive.

Deep learning models use non-linear transformations on input data in a series of specified layers in order to best fit outcomes [146]. The benefits of deep learning models are that perform well on highly complex and high-dimensional datasets and the presence of hidden layers reduce the need for feature selection. However, deep learning models require large datasets to perform

well, can have lower performance than ensemble tree models, are computationally expensive, and are challenging for non-experts to tune.

The Stacked Ensemble method within H2O uncovers the best combination of machine learning frameworks that feed into one another. Namely, the Stacked Ensemble pulls from generalized linear models, distributed random forest, gradient boosting machine, deep learning, XGBoost, and Naïve Bayes in order to uncover the optimal combination of prediction models. Five-fold cross validation was conducted to allow for greater generalizability. Models were tested on the top 5, 10, 25, and 50 features for predicting both treatment response and symptom remission.

7.2 Results

7.2.1 Feature selection

Mutual information was calculated between each feature and the outcome variable. Table 20 holds the top 20 features according to greatest mutual information score and after filtering based on a Pearson's correlation coefficient threshold less than 0.85. "Myalgia" and "Myalgia and myositis, unspecified" were found to have a Pearson's correlation coefficient of 0.99. In addition, "Long term current use of anticoagulant" and "Long term current use of anticoagulant therapy" were found to be highly correlated with a Pearson's correlation coefficient of 0.97. "Spinal stenosis" and "Spinal stenosis, lumbar" also had a correlation coefficient above the 0.85 threshold at 0.86, along with "Spinal stenosis lumbar region without neurogenic claudication" which had a correlation coefficient with "Spinal stenosis" of 0.86. Correlation coefficients between the final top 5, 10, and 25 features for predicting treatment response (Figure 21) and predicting symptom

remission (Figure 22) were less than 0.85. Figures 21 and 22 are square matrices conveying Pearson correlation coefficients between the top 25 features for predicting treatment response and symptom remission, respectively.

Table 20 Mutual information scores for top 25 features associated with both treatment response and symptom remission outcome variables.

Feature	Mutual information score (Treatment response)	Mutual information score (Symptom remission)
Time between first MDD ICD-9/-10 code and first antidepressant	1.20×10^{-1}	N/A
Weight	1.16×10^{-2}	1.43×10^{-1}
Insomnia	1.08×10^{-2}	N/A
Age	1.08×10^{-2}	4.26×10^{-2}
Routine infant or child health check	6.40×10^{-3}	N/A
Insomnia, unspecified	6.03×10^{-3}	N/A
Fibromyalgia	5.99×10^{-3}	3.47×10^{-2}
Encounter for routine child health exam without abnormal findings	5.81×10^{-3}	N/A
BMI	5.74×10^{-3}	3.23×10^{-2}
Chronic pain	5.35×10^{-3}	N/A
Myalgia and myositis, unspecified	5.33×10^{-3}	2.94×10^{-2}
Need for other specified prophylactic vaccination against single bacterial disease	5.27×10^{-3}	N/A
Systolic blood pressure	4.84×10^{-3}	3.43×10^{-2}
Long term current use of anticoagulant	4.83×10^{-3}	6.07×10^{-2}
Encounter for general adult medical exam without abnormal findings	4.48×10^{-3}	N/A
Myalgia	4.40×10^{-3}	N/A
Primary insomnia	4.19×10^{-3}	N/A
Migraine	4.18×10^{-3}	N/A
Encounter for immunization	4.01×10^{-3}	N/A
Insomnia, unspecified type	3.82×10^{-3}	N/A
Diastolic blood pressure	3.80×10^{-3}	N/A
Pain	3.76×10^{-3}	2.72×10^{-2}
Dietary counseling and surveillance	3.69×10^{-3}	N/A
Anxiety	3.65×10^{-3}	3.30×10^{-2}
Essential hypertension	3.59×10^{-3}	3.98×10^{-2}
Chronic atrial fibrillation	N/A	3.88×10^{-2}
Hyperlipidemia	N/A	3.61×10^{-2}
Spinal stenosis	N/A	3.41×10^{-2}
Pure hypercholesterolemia	N/A	3.31×10^{-2}
Hypothyroidism	N/A	3.29×10^{-2}
Other unspecified hyperlipidemia	N/A	2.97×10^{-2}
Hypertension	N/A	2.97×10^{-2}
Unspecified essential hypertension	N/A	2.90×10^{-2}
Mixed hyperlipidemia	N/A	2.89×10^{-2}
Depression	N/A	2.84×10^{-2}

Table 20 (Continued)

Feature	Mutual information score (Treatment response)	Mutual information score (Symptom remission)
B12 deficiency	N/A	2.76×10^{-2}
Benign essential hypertension	N/A	2.65×10^{-2}
Coronary artery disease involving native coronary artery of native heart without angina pectoris	N/A	2.47×10^{-2}
Depressive disorder	N/A	2.44×10^{-2}
Unspecified hypothyroidism	N/A	2.43×10^{-2}

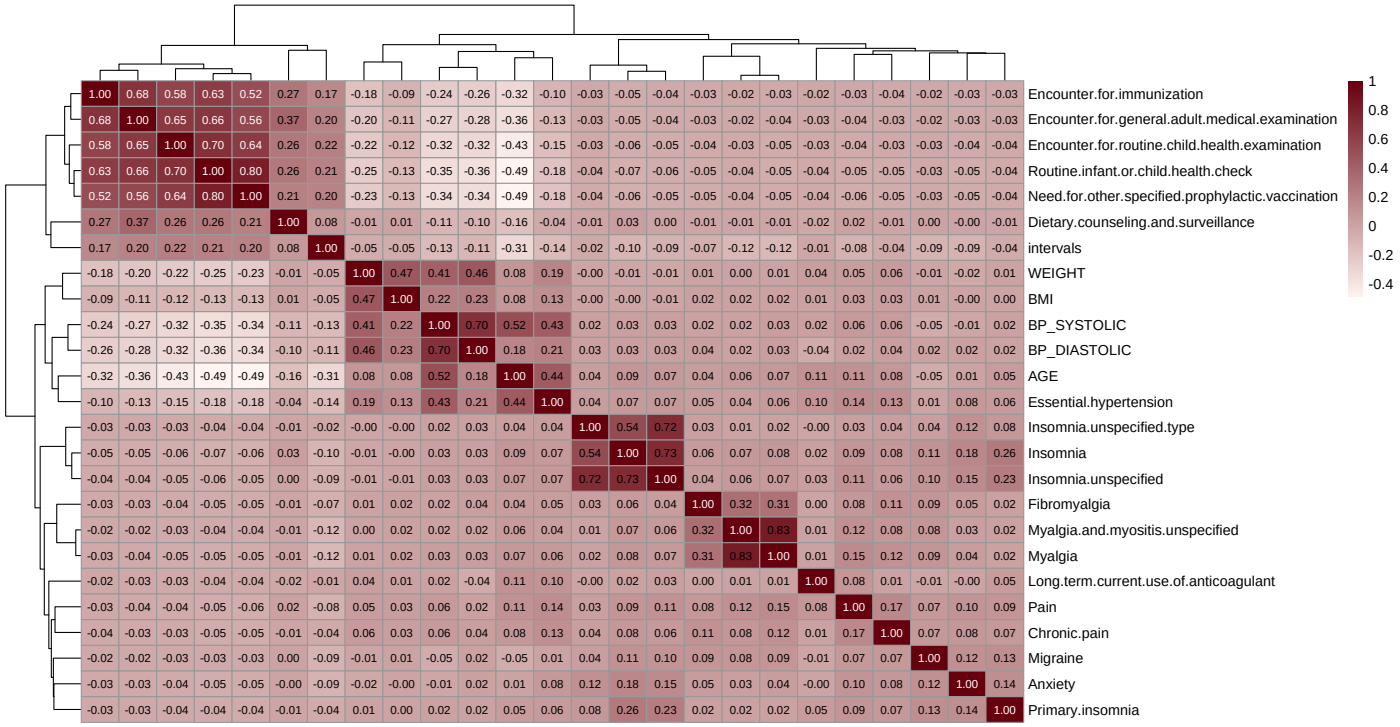


Figure 21 Pearson correlation coefficients between the top 25 features for predicting treatment response.

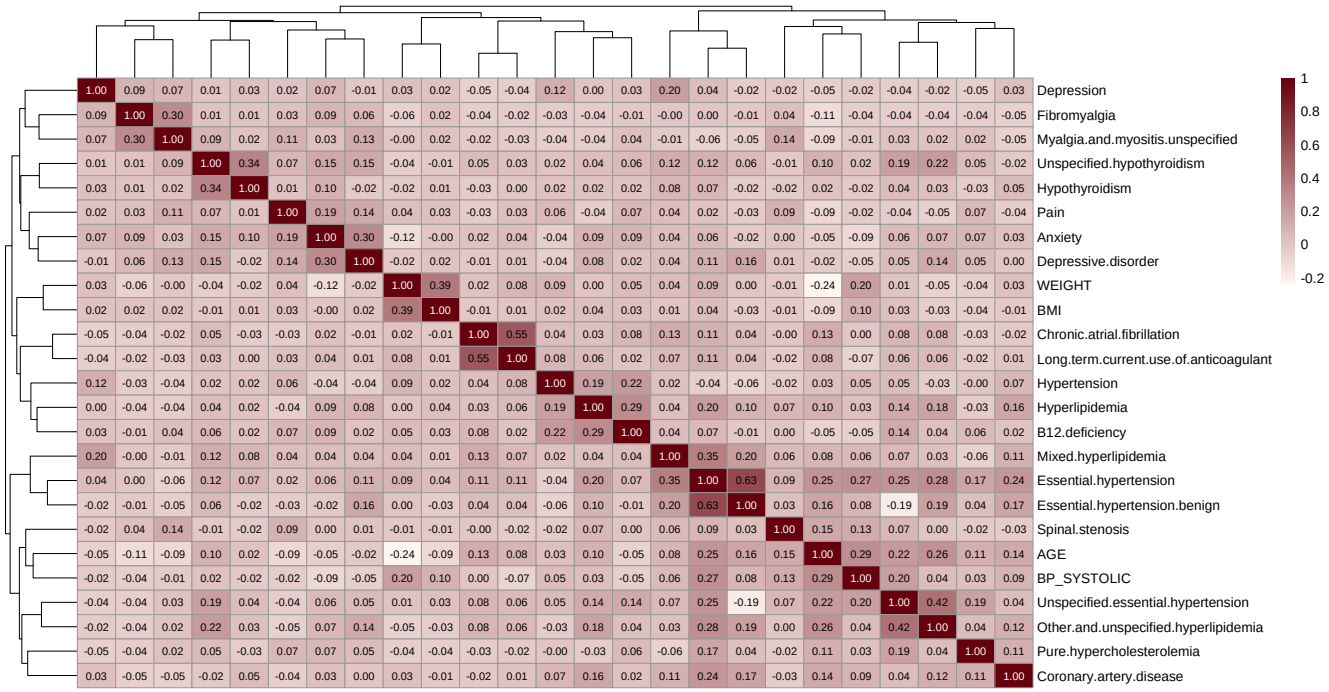


Figure 22 Pearson correlation coefficients between the top 25 features for predicting symptom remission.

7.2.2 Logistic regression

The logistic regression models for the a priori model for treatment response (Table 21) had an Akaike information criterion (AIC) of 19,520. The logistic regression models created on the top five features for predicting treatment response had an AIC of 17,314 (Table 22). Logistic regression models created using the top 10 features for predicting treatment response had an AIC of 17,268 (Table 23), while the models created using the top 25 features (Table 24) and top 50 features for predicting treatment response both had an AIC of 17,226. In order to minimize the probability of information loss, the treatment response model with the lowest AIC and the fewest number of features was deemed the best model, which is the model created using 25 features. The highest performing logistic regression model for predicting treatment response was the model using the top 50 features, which had an accuracy of 77.94% and an F1 score of 87.56%.

Table 21 Odds ratios of coefficients in the a priori model for treatment response prediction.

Features	Estimate	Standard Error	P-value
Intercept	0.16	6.13×10^{-2}	< 0.0001
Age	0.50	9.0×10^{-4}	< 0.0001
Gender Male	0.47	4.06×10^{-2}	0.00206
Race American Indian	0.48	5.05×10^{-1}	0.90
Race Asian Pacific Islander	0.54	1.79×10^{-1}	0.41
Race Black	0.50	8.19×10^{-2}	0.83
Race Unknown	0.49	1.52×10^{-1}	0.83
Ethnicity Hispanic	0.54	1.75×10^{-1}	0.30
Ethnicity Unknown	0.54	8.85×10^{-2}	0.10
BMI	0.50	1.02×10^{-3}	0.04

Table 22 Odds ratios for coefficients in the logistic regression model for treatment response prediction using top five features.

Features	Estimate	Standard Error	P-value
Intercept	0.23	0.10	< 0.0001
Age	0.50	1.12×10^{-3}	0.11
Insomnia	0.51	5.30×10^{-3}	< 0.0001
Routine infant or child health check	0.43	3.35×10^{-2}	< 0.0001
Weight	0.50	3.77×10^{-4}	0.27
Time between first MDD ICD-9/-10 code and first antidepressant	0.50	2.17×10^{-5}	< 0.0001

Table 23 Odds ratios of coefficients in logistic regression model for treatment response prediction using top ten features.

Features	Estimate	Standard Error	P-value
Intercept	0.24	0.10	< 0.0001
Age	0.50	1.13×10^{-3}	0.34
Fibromyalgia	0.51	3.98×10^{-3}	< 0.0001
Insomnia	0.51	8.96×10^{-3}	< 0.0001
Insomnia, unspecified	0.49	1.25×10^{-2}	0.01
Routine infant or child health check	0.45	4.06×10^{-2}	< 0.0001
Routine infant or child health check without abnormal findings	0.43	7.20×10^{-2}	2.10×10^{-4}
Chronic pain	0.50	2.42×10^{-3}	0.03
Weight	0.50	4.15×10^{-4}	0.04
BMI	0.50	1.23×10^{-3}	0.16
Time between first MDD ICD-9/-10 code and first antidepressant	0.50	2.17×10^{-5}	< 0.0001

Table 24 Odds ratios of coefficients in logistic regression model for treatment response prediction using top 25 features.

Features	Estimate	Standard Error	P-value
Intercept	0.21	0.28	< 0.0001
Age	0.50	1.42×10^{-3}	0.06
Long term current use of anticoagulant therapy	0.49	3.40×10^{-2}	0.12
Fibromyalgia	0.50	4.22×10^{-3}	2.0×10^{-3}
Myalgia and myositis	0.49	5.59×10^{-2}	0.67
Encounter for immunization	0.48	3.63×10^{-2}	0.04
Primary insomnia	0.50	7.61×10^{-3}	0.05
Insomnia	0.51	9.02×10^{-3}	< 0.0001
Myalgia	0.50	2.21×10^{-2}	0.43
Insomnia, unspecified type	0.51	1.76×10^{-2}	0.02
Insomnia, unspecified	0.49	1.47×10^{-2}	< 0.0001
Routine infant or child health check	0.49	5.35×10^{-2}	0.33
Dietary counseling and surveillance	0.50	2.12×10^{-2}	0.93
Pain	0.50	3.70×10^{-3}	0.19
Migraine	0.50	4.35×10^{-3}	< 0.0001
Routine infant or child health check without abnormal findings	0.46	7.50×10^{-2}	0.03
General adult medical exam without abnormal findings	0.47	8.32×10^{-2}	0.13

Table 24 (Continued)

Features	Estimate	Standard Error	P-value
Need for prophylactic vaccination against single bacterial disease	0.43	1.39×10^{-1}	0.04
Chronic pain	0.50	2.52×10^{-3}	0.2
Long term current use of anticoagulant	0.51	3.38×10^{-2}	0.11
Myalgia and myositis	0.51	5.69×10^{-2}	0.42
Weight	0.50	4.56×10^{-4}	0.03
BMI	0.50	1.24×10^{-3}	0.13
Systolic blood pressure	0.50	3.20×10^{-3}	0.05
Diastolic blood pressure	0.50	4.78×10^{-3}	0.02
Time between first MDD ICD-9/-10 code and first antidepressant	0.50	2.21×10^{-5}	< 0.0001

The logistic regression models for the a priori model for symptom remission (Table 25) had an Akaike information criterion (AIC) of 1,102. The logistic regression model created on the top five features for predicting symptom remission had an AIC value of 1,371.4 (Table 26). Logistic regression model using the top 10 features for predicting symptom remission had an AIC of 1,363.3 (Table 27), while the model created using the top 25 features (Table 28) had an AIC of 1,339.3. The minimum AIC came from the logistic regression model using the top 25 features to predict symptom remission. The prediction accuracy for the logistic regression models for symptom remission using the top five features was 55.22%, top ten features was 62.19%, top 25 features was 64.67%, and top 50 features was 56.72% (Table 29). Therefore, the prediction

accuracy was highest for the logistic regression model using the top 25 features for predicting symptom remission. The highest performing logistic regression model for predicting symptom remission was the model using the top 50 features, which had an accuracy of 55.22% and an F1 score of 39.19%.

Table 25 Odds ratios of coefficients in the a priori model for symptom remission prediction.

Features	Estimate	Standard Error	P-value
Intercept	0.39	4.24×10^{-1}	0.29
Age	0.50	5.0×10^{-3}	5.13×10^{-3}
Gender Male	0.60	1.60×10^{-1}	1.04×10^{-2}
Race American Indian	0.45	1.42	0.88
Race Asian Pacific Islander	0.38	1.02	0.68
Race Black	0.35	2.81×10^{-1}	0.03
Race Unknown	0.47	1.05	0.89
Ethnicity Hispanic	0.66	1.24	0.59
Ethnicity Unknown	0.52	3.72×10^{-1}	0.85
BMI	0.50	6.16×10^{-3}	0.15

Table 26 Odds ratios of coefficients in logistic regression model for symptom remission prediction using top five features.

Features	Estimate	Standard Error	P-value
Intercept	0.30	0.46	0.07
Age	0.50	4.78×10^{-3}	2.11×10^{-3}
Essential hypertension	0.50	4.68×10^{-3}	0.22
Chronic atrial fibrillation	0.50	3.91×10^{-3}	0.47
Long term current use of anticoagulant	0.50	2.50×10^{-3}	0.84
Weight	0.50	1.30×10^{-3}	0.96

Table 27 Odds ratios of coefficients in logistic regression model for symptom remission prediction using top ten features.

Features	Estimate	Standard Error	P-value
Intercept	0.48	9.28×10^{-1}	0.94
Age	0.50	5.15×10^{-3}	4.41×10^{-3}
Essential hypertension	0.50	5.02×10^{-3}	0.34
Pure hypercholesterolemia	0.50	1.11×10^{-2}	6.75×10^{-2}
Hyperlipidemia	0.50	9.32×10^{-3}	0.07
Fibromyalgia	0.49	1.20×10^{-2}	1.97×10^{-2}
Chronic atrial fibrillation	0.50	3.90×10^{-3}	0.43
Spinal stenosis	0.50	6.87×10^{-3}	0.08
Long term current use of anticoagulant	0.50	2.55×10^{-3}	0.98
Weight	0.50	1.37×10^{-3}	0.89
Systolic blood pressure	0.50	7.74×10^{-3}	0.41

Table 28 Odds ratios of coefficients in logistic regression model for symptom remission prediction using top 25 features.

Features	Estimate	Standard Error	P-value
Intercept	0.67	0.99	0.46
Age	0.50	5.47×10^{-3}	0.16
Essential hypertension	0.50	7.78×10^{-3}	0.49
Unspecified essential hypertension	0.50	9.09×10^{-3}	0.89
Anxiety	0.50	6.86×10^{-3}	0.05
Other and unspecified hyperlipidemia	0.51	9.67×10^{-3}	0.01
Mixed hyperlipidemia	0.51	9.68×10^{-3}	7.60×10^{-3}
Benign essential hypertension	0.50	9.30×10^{-3}	0.64
Depression	0.50	1.27×10^{-2}	0.39
Unspecified hypothyroidism	0.50	1.36×10^{-2}	0.91
Pure hypercholesterolemia	0.51	1.17×10^{-2}	0.01
Hyperlipidemia	0.50	1.05×10^{-2}	0.08
Fibromyalgia	0.50	1.24×10^{-2}	0.40
B12 deficiency	0.50	1.13×10^{-2}	0.93
Hypertension	0.50	6.37×10^{-3}	0.26
Coronary artery disease involving native coronary artery of native heart without angina pectoris	0.51	1.55×10^{-2}	0.03
Chronic atrial fibrillation	0.50	4.04×10^{-3}	0.67
Unspecified myalgia and myositis	0.49	1.70×10^{-2}	0.03

Table 28 (Continued)

Features	Estimate	Standard Error	P-value
Depressive disorder	0.49	1.63×10^{-2}	7.64×10^{-3}
Hypothyroidism	0.50	1.32×10^{-2}	0.42
Pain	0.49	1.11×10^{-2}	0.04
Spinal stenosis	0.50	7.22×10^{-3}	0.31
Long term current use of anticoagulant	0.50	2.67×10^{-3}	0.79
Weight	0.50	1.87×10^{-3}	0.21
BMI	0.49	9.56×10^{-3}	0.03
Systolic blood pressure	0.50	8.17×10^{-3}	0.46

Table 29 Logistic regression model performance

Outcome predicted	No. patients	No. cases	Model feature space	TPR (%)	1-FPR (%)	PPV (%)	NPV (%)	FPR (%)	FDR (%)	FNR (%)	F1 score (%)	Accuracy (%)
Treatment response	16,482	3,746 (22.72%)	Top 5 features	77.84	25.00	99.42	0.68	75.00	0.58	22.16	87.31	77.52
			Top 10 features	77.37	35.29	99.57	0.80	64.71	0.43	22.63	87.08	77.15
			Top 25 features	78.33	34.21	99.03	1.81	65.79	0.97	21.67	87.47	77.82
			Top 50 features	78.35	33.33	99.22	1.39	66.67	0.78	21.65	87.56	77.94
Symptom remission	1,008	576 (57.14%)	Top 5 features	63.16	55.49	12.90	93.52	44.51	87.10	36.84	21.43	56.22
			Top 10 features	55.26	55.83	22.58	84.26	44.17	77.42	44.74	32.06	55.72
			Top 25 features	52.83	56.08	30.11	76.85	43.92	69.89	47.17	38.36	55.22
			Top 50 features	42.65	61.65	36.25	67.77	38.35	63.75	57.35	39.19	55.22

7.2.3 Random forest

A random forest model was constructed using the top 5, 10, 25, and 50 features for predicting both treatment response and symptom remission (Table 30). Five-fold cross validation was conducted for each model. The top performing model for predicting treatment response used the top 25 features and had an accuracy of 77.21% and an F1 score of 87.07%. The greatest F1 score for predicting symptom remission was 47.85% using the top 50 features, however the greatest accuracy of 59.20% came from using the top 25 features.

Table 30 Random forest model performance

Outcome predicted	No. patients	No. cases	Model feature space	No. of trees	No. of vars tried at split	TPR (%)	1-FPR (%)	PPV (%)	NPV (%)	FPR (%)	FDR (%)	FNR (%)	F1 score (%)	Accuracy (%)
Treatment response	16,482	3,746 (22.72%)	Top 5 features	500	2	77.33	24.51	96.98	3.34	75.49	3.02	22.67	86.05	75.70
			Top 10 features	500	3	77.33	26.39	97.92	2.53	73.61	2.08	22.67	86.41	76.21
			Top 25 features	500	5	77.40	56.67	99.49	2.25	43.33	0.51	22.60	87.07	77.21
			Top 50 features	500	7	76.96	42.86	99.37	1.57	57.14	0.63	23.04	86.74	76.67
Symptom remission	1,008	576 (57.14%)	Top 5 features	500	2	46.43	57.93	29.89	73.68	42.07	70.11	53.57	36.36	54.73
			Top 10 features	500	3	51.67	53.90	32.29	72.38	46.10	67.71	48.33	39.74	53.23
			Top 25 features	500	5	50.00	62.43	31.71	78.15	37.58	68.29	50.00	38.81	59.20
			Top 50 features	500	7	55.71	58.78	41.94	71.30	41.22	58.06	44.29	47.85	57.71

7.2.4 Auto ML

The best performing prediction model for treatment response within the analysis cohort used the top 50 features for predicting treatment response and consisted of a stacked ensemble incorporating 1 deep learning frameworks, 1 distributed random forest, 1 gradient boosting machines, 1 generalized linear model, and 1 XGBoost (Table 31). The accuracy was 60.38%, and the F1 score was 70.36%. Given that the outcome variables were imbalanced, the F1 score provides the best prediction metric to be evaluated. The best performing prediction model for predicting symptom remission used the top 25 features for predicting symptom remission and was a generalized linear model. The accuracy for this model was 68.16% and the F1 score was 33.33%.

Table 31 Auto ML model performance

Outcome predicted	No. cases / No. patients, (%)	Model feature space	Best prediction model	Model architecture	AUC (%)	AUCPR (%)	TPR (%)	1-FPR (%)	PPV (%)	NPV (%)	FPR (%)	FDR (%)	FNR (%)	F1 score (%)	Accuracy (%)
Treatment response	3,746 / 16,482 (22.72)	Top 5 features	Stacked ensemble	1 deep learning, 2 drf, 1 gbm, 1 glm, 1 xgboost	62.61	33.39	83.95	29.66	47.09	71.25	70.34	52.91	16.05	60.34	52.85
		Top 10 features	Stacked ensemble	1 deep learning, 2 drf, 7 gbm, 1 glm, 5 xgboost	61.09	29.60	85.21	25.88	34.26	79.43	74.12	65.74	14.79	48.87	44.39
		Top 25 features	Stacked ensemble	1 deep learning, 2 drf, 1 gbm, 1 glm, 1 xgboost	61.83	30.28	84.96	26.34	41.34	74.14	73.66	58.66	15.04	55.62	48.57
		Top 50 features	Stacked ensemble	1 deep learning, 1 drf, 1 gbm, 1 glm, 1 xgboost	63.17	31.44	84.28	30.20	60.38	60.36	69.80	39.62	15.72	70.36	60.38
Symptom remission	576 / 1,008 (57.14)	Top 5 features	gbm	N/A	49.48	49.26	85.71	52.58	6.12	99.03	47.42	93.88	14.29	11.43	53.73
		Top 10 features	glm	N/A	62.35	71.59	100	60.50	1.25	100	39.50	98.75	0.00	2.47	60.70
		Top 25 features	glm	N/A	64.96	70.73	84.21	66.48	20.78	97.58	33.52	79.22	15.79	33.33	68.16
		Top 50 features	Stacked ensemble	1 deep learning, 2 drf, 1 gbm, 1 glm, 1 xgboost	62.15	64.41	75.00	61.33	17.65	95.69	38.67	82.35	25.00	28.57	62.69

drf: distributed random forest, gbm: gradient boosting machine, glm: generalized linear model, xgboost: eXtreme gradient boosting.

7.3 Discussion

Many of the top features were related to pain, namely, “fibromyalgia”, “chronic pain”, “myalgia and myositis”, “pain”, and “migraine”. In addition, many features were related to lack of sleep: “insomnia”, “insomnia, unspecified”, “primary insomnia”, and “insomnia unspecified type”. Another observed pattern within the top 25 features were physiological, with “weight”, “age”, “BMI”, “systolic blood pressure”, and “diastolic blood pressure”. The relationship between pain and depression has been reported in depth on previously, in addition to psychological symptoms like disturbed sleep [147]. The presence of fibromyalgia and pain as significant diagnoses associated with depression have also been replicated in other studies as well [137]. Investigators have sought to parse a causal relationship between pain and depression with mixed conclusions [148].

Another feature of depression that is widely reported in literature is the contributions of structural risk factors from gender and race in connection to MDD for patients [149–154]. However, gender and race were not within the top 50 features by mutual information score for patients within the analysis cohort, and therefore were not strongly related to treatment response and symptom remission prediction. The mutual information score for gender and treatment response was 4.55×10^{-4}, while the mutual information score for race and treatment response was 8.12×10^{-5}. These mutual information scores were one and two orders of magnitude less than the 50^{th} greatest mutual information score for treatment response (2.41×10^{-3}). For symptom remission, the mutual information score for gender and symptom remission was 4.31×10^{-3}, and the mutual information score for race and symptom remission was 4.95×10^{-3}. These mutual

information scores were one order of magnitude less than the 50^{th} greatest mutual information score for symptom remission (1.93×10^{-2}). The literature supporting gender and race influencing MDD for patients interestingly did not replicate in our study.

Despite this lack of replication, a strength of our study was the high performing logistic regression models. Logistic regression models are beneficial for model interpretability, and therefore have a greater likelihood of being adopted in clinical practice based on the intuitive equation. Model interpretability is paramount to physician trust in models and therefore in model adoption in clinical practice [155]. However, interpretability notwithstanding, logistic regression models had similar performance to the random forest and ensemble methods. The highest performing logistic regression model for predicting treatment response had an accuracy of 77.94%, while the highest performing random forest for predicting treatment response had an accuracy of 77.21%, and the highest performing ensemble method had an accuracy of 60.38%. These metrics for predicting treatment response are greater than those in previously published literature by Chekroud et al. (2017) [93], though these models did not perform as well as deep learning models published by Lin et al. (2018) that had a sensitivity of 75.46% and specificity of 69.22% [97].

For predicting symptom remission, the highest performing logistic regression model accuracy was of 55.22%, while the highest performing random forest had an accuracy of 59.20%. Likewise, the highest performing ensemble method had an accuracy of 68.16% in predicting symptom remission. These performance metrics are not as high as models from Iniesta et al. (2018) [156] using an elastic net logistic model or Lin et al. (2018) [97] using a multilayer feedforward neural network with 3 hidden layers to predict symptom remission.

Despite the room for improvement in these models' performance metrics, due to the fact that models in this study performed similarly, a logistic regression model might be the most

preferable to aid clinicians in prescribing decisions about which patients are likely to respond to first-line therapies due to factors concerning model interpretability.

In order to improve on these performance metrics, future directions might concern improving upon feature selection methods. While correlation between features was taken into consideration, and feature ranking was considered beneficial for its reduced computational requirements, future directions could include using methods that incorporate the classification task in narrowing down the feature set as opposed to feature ranking methods that might not select the optimal feature subset [157]. In addition to feature selection methods, future directions might concern expanding the feature space to include the individual antidepressant or antidepressant class that a patient was prescribed as first-line therapy, or models could be created specifically to predict treatment response or symptom remission to a particular antidepressant drug or drug class.

In addition, the incorporation of PGx data might also help to improve performance metrics. The other models cited predicting treatment response and symptom remission used genetic variant data as part of the feature space. Pharmacogenomic data to characterize patients' metabolizer status along with the clinical EHR data might help to explain variability in treatment response and resolution of depression symptoms and therefore improve MDD treatment outcome prediction.

8.0 Development of a Reproducible Machine Learning Pipeline

8.1 Methods

The EHR database was provided in CSV format. All programming was conducted in the R Statistical Programming language (version 4.0.4) [158]. All code was written by Lauren Rost, using packages lubridate [159], dplyr [160], RColorBrewer [161], pheatmap [162], randomForest [163], markovchain [100], and h2o [142]. All algorithms were run on the laptop environment MacBook Pro with a 2.5 GHz Quad-Core Intel Core i7 processor and 16 GB of RAM, running the 64-bit macOS Big Sur operating system. Algorithmic methods were based on literature review of analyses for major depressive disorder clinical analyses.

A reproducible data analysis and machine learning pipeline was created in order to allow for the study to be re-run as additional patients are enrolled in the Pitt + Me Discovery cohort. The pipeline is made up of 14 R code files, where each script feeds into one another (Table 32). The input file for the first file is used to create an output file, which becomes the next input file for the next script. This pipeline could be used for other diagnoses as well in order to map out the sequence of medications that patients were prescribed and to draw conclusions about prescribing patterns.

The first R script, 1_InclusionCodes.R checks for the presence of the ICD-9/-10 codes for MDD. This script accepts the table in the database containing diagnosis information and outputs a file containing ICD-9/-10 inclusion codes for MDD as the column names and has patient STUDY_IDs as the row names. The matrix is populated by the number of ICD-9/-10 codes present for each patient. The subsequent script in the pipeline 2_ExclusionCodes.R operates similarly,

except that it checks for ICD-9/-10 codes that should be excluded based on associated treatment patterns, as exhibited in schizophrenia, bipolar disorder, autism, etc.

The next R script in the pipeline 3_PullDepressionCodeDates.R accepts the output of 1_InclusionCodes.R and performs a more directed search for the ICD-9/-10 code based on whether the code is present for that patient. This directed search allows for greater computational efficiency in not searching through all patients for all codes.

After, the 4_TwoThirty180.R script checks for whether patients pass the 2/30/180 rule from eMERGE. Therefore, this script checks when an ICD-9/-10 code for MDD appears in a patient's record, and determines whether there are at least two ICD-9/-10 codes present in the EHR, and whether there are at least two ICD-9/-10 codes that are at least 30 days apart and no greater than 180 days apart. With this information, the 5_PullAntidepressants.R script then has a refined list of STUDY_IDs that pass the 2/30/180 rule, and the script can capture the antidepressants that patients were prescribed. 6_PullAntidepressantDates.R works very similarly to 5_PullAntidepressants.R in that the script pulls the date for which the antidepressant was prescribed for each patient. Both scripts output a file that has each patients' STUDY_ID as row names, with antidepressants or dates going across the columns. Afterwards, 7_OrderedCodesandDates.R uses the matrices of antidepressants and antidepressant prescribing dates to put the antidepressants prescribed in chronological order for each patient.

From there, 8_OutcomeLabelling.R uses the ordered antidepressants for each patient and the antidepressant prescription dates to determine whether the patient switched or continued the drug after at least eight weeks. Additional scripts then created additional features for the machine learning models, like 9_FeatureEngineer_DateCorrespondence_AntidepressantsandCodes.R which calculates the time between when an antidepressant was prescribed and when the first MDD

ICD-9/-10 code was present in the electronic health record. The 9_FeatureEngineer_Questionnaire.v4 script uses the depression severity score data to plot PHQ scores over time, and to label symptom remission and treatment response based on score thresholds and relative score changes for patients at each of the PHQ scores on file.

Then, 10_MarkovModel.R creates a Markov Chain Model using the sequence of antidepressants prescribed for each patient to examine transition probabilities between antidepressants. This script also creates Markov Chain Model figures within this study. The 13th R script calculates antidepressant sequence pattern statistics and plots sequence patterns. Finally, the 14th script runs machine learning models on subsets of patients; all patients prescribed antidepressants and surviving the 2/30/180 rule, patients receiving behavioral therapy, patients with PGx data, and patients with PHQ scores on file.

Table 32 Reproducible data analysis and machine learning pipeline

Step	Code name	Function
1	1_InclusionCodes.R	Checks for presence of inclusion ICD-9/10 codes
2	2_ExclusionCodes.R	Checks for presence of exclusion ICD-9/10 codes
3	3_PullDepressionCodeDates.R	Finds the dates associated with ICD codes
4	4_TwoThirty180.R	Checks whether patients pass the 2/30/180 rule
5	5_PullAntidepressants.R	Pulls antidepressant sequences
6	5_PullDiagnoses.R	Pulls diagnoses associated with patients' EHR
6	6_PullAntidepressantDates.R	Pulls dates of antidepressants
7	7_OrderedICDCodesandDates.R	Orders ICD codes according to when they were documented in the EHR
8	7_OrderedAntidepressantsandDates.R	Puts antidepressants in sequential order according to when they were documented
9	8_OutcomeLabelling.R	Labels whether the patients' prescribed antidepressant was continued within 12 weeks
10	9_FeatureSelection.R	Calculates mutual information scores between each feature and the outcome variable

8.2 Results

The reproducible pipeline can be accessed at https://github.com/laurenrost/ReproduciblePipeline. The pipeline consisted of 10 steps, where the output of each step in the pipeline fed into the following step of the pipeline. The code was designed to be amenable to applications outside of MDD diagnosis and an antidepressant medication list. Any diagnosis or medication list of interest could be input and one could examine the sequential nature of drugs prescribed or diagnosis codes assigned. This sequential order could

then be annotated to reflect treatment response definitions or modelled to uncover the probabilities of switching between drugs.

8.3 Discussion

This reproducible pipeline can be applied to additional diagnosis lists outside of MDD and medication lists besides antidepressants. The code is thoroughly annotated to allow for greater understanding and generalizability to novel applications. This pipeline will ideally contribute to greater reproducibility and extend findings from EHR data. It has been noted that many EHR studies are not reproducible due to the fact that preprocessing, cleaning, phenotyping and analysis methods are not shared [164]. Denaxas et al. (2017) recommended producing generic functions that conduct data cleaning and preprocessing, creating adaptable functions for defining outcomes, exposures, and covariates, constructing modules for study population definitions and subsets, hosting annotated machine-readable EHR phenotyping algorithms, and using both logical operators and programming commands for literate programming, which this pipeline can be found in adherence. Future directions could involve migrating this pipeline onto a virtual machine or a Docker container to allow for potentially more user-friendly reproducibility.

In addition, this pipeline was made reproducible not only for outside applications, but also for it to be re-run on its original application, for MDD and antidepressants, with the addition of PGx data. PGx data will add additional features to inform prediction models.

9.0 Conclusions and Future Directions

We created a reproducible data analysis and machine learning pipeline that intakes EHR data and a medication list, to document the history of medications that patients were prescribed. These sequences of prescriptions are then ordered according to their prescription date and subsequently, transition probabilities between antidepressant prescriptions are calculated and conveyed in Markov chain models. The most frequently ordered initial antidepressants are conveyed, along with the most frequent two-drug medication order sequences and three-drug medication order sequences. Once the medication order results are conveyed, subsets of the total analysis cohort are analyzed to predict treatment response and symptom remission using logistic regression models, random forest models, and an automated machine learning package that uncovers optimal ensemble prediction models. It is important to note that the models created through this pipeline, both the Markov chain models and statistical models, are not causal. Treatment patterns are not entirely random, and therefore, randomized control trials are necessary to address causality and treatment effectiveness.

However, this reproducible data analysis and machine learning pipeline does offer a way for modeling prescribing patterns and pharmacologic treatment response prediction. This pipeline especially allows biomedical informatics data science to be more accessible to individuals with less programming experience. Individuals with motivations to analyze other diagnoses and medications sequences for patients within EHR data can also use this pipeline. In addition, the machine learning portion of this pipeline also allows individuals with less expertise in machine learning to run prediction analyses.

Performance of the treatment response and symptom remission prediction models were not ostensibly favorable enough to be adopted into a clinical workflow to assist with prescribing. However, the models could be used to aid in prescribing decisions of whether first-line therapies should be reconsidered. In order to demonstrate the generalizability and portability of this model, the pipeline should be run on data from another hospital. The importance of EHR machine learning models generalizing across hospitals has been emphasized heavily [165,166].

Another interesting finding of this real-world EHR analysis was the duration of time to follow-up appointment after original antidepressant prescription for some patients. Given the importance of early assessment of treatment response in terms of both efficacy and tolerability when initiating antidepressant therapy, opportunities to improve the timeliness of follow-up could lead to improved medication adherence and better treatment outcomes for patients with MDD.

Future directions include the implementation of state-of-the-art machine learning models using EHR data, namely a stacked denoising autoencoder to predict symptom remission and treatment response as binary outcomes. Other neural network architectures that could be implemented or adapted from similar work are a variational autoencoder (AE) model [167], long-short term memory, convolutional neural network, recurrent neural network, transformer (with attention between encoder and decoder), ELMO (based on LSTM neurons), BERT (feed-forward architecture built off of transformer units for sequence representation) [168], BEHRT (transformer for EHR) [169], RETAIN (attention-based recurrent neural network, or two RNNs in parallel) [170].

This work is novel in examining electronic health record data from major depressive disorder patients and seeks to join PGx data in order to inform symptom remission and treatment response prediction. The PGx data is high-dimensional be it from a panel of 4,626 markers within

1,191 genes. In addition, the data analysis pipeline structure allows the data analysis to be re-run as additional patients are enrolled.

This work also adds to the field in that there is a large patient population examined. This study included many antidepressant classes unlike many studies that only observe SSRIs. This work is also novel in contributing to clinical outcomes metrics including adhering to best practices and following clinical recommendations and guidelines. In addition, this work contributes to research for new genotype-phenotype discovery and validation. The value, interest, and infrastructure for this pipeline and return of results exists; and the reproducible machine learning pipeline is robust to be run again and again as more PGx data become available to eventually draw exciting conclusions and predictions for treatment of MDD at UPMC.

Despite the exciting potential, this study has many drawbacks and limitations. Namely, there are assumptions made surrounding the labels of whether patients experienced antidepressant treatment response or symptom remission from an antidepressant or not. Treatment response is proxied by whether a subsequent prescription was continued within a twelve week time period. There are many clinical factors that affect patient experience with MDD and its treatment that are not captured in our models. Namely, social determinants of health and other patient-level factors including stressful life events, cultural and religious beliefs, perceptions, socioeconomic status, degree of family support, and access to care which are not captured in the EHR database. To temper this, a future direction might be to extract additional information about the patient experience from the clinical note in order to better inform the treatment prediction task. Another future direction to incorporate additional data types to inform prediction models could be the inclusion of personal monitoring data, especially due to the significant features concerning physiological factors like BMI, blood pressure, and an ICD-9/-10 code for dietary counseling and surveillance.

For patients that we do have additional information about their experience with the antidepressant in the form of depression severity rating scale data, there are still limitations in the quality of this data because depression severity rating scales are psychometric, and therefore are not a gold standard quantitative measure for patients' experience with depression including level of functioning and quality of life.

This study also has inherent limitations due to its observational nature. This observational retrospective cohort will likely not generalize well to other cohorts for temporal and recruitment factors. The fact that this data came from only one health system also limits the generalizability of the model and findings to other clinical care settings as well. In addition, due to the fact that this was not a prospective study or randomized control trial, prescription insurance coverage might play a role in the medications that were prescribed, and therefore might have led to gaps and biases that could undermine results. There also might be gaps and biases in the data due to discontinuations in patient records, for example if patients switched from UPMC medical providers resulting in a gap in patient data. This limitation could be ameliorated by analyzing health insurance claims data which more comprehensively capture patients' encounters with medical providers.

In this study, a reproducible analysis pipeline was created to model antidepressant treatment sequences and inform prediction models for MDD treatment outcomes. This pipeline was designed to be re-run using PGx data to demonstrate whether models experience improved prediction performance using PGx metabolizer phenotype features. In addition, this reproducible pipeline was designed to be re-used on other diagnoses and medication lists, and will hopefully serve to facilitate and contribute to future EHR research.

Appendix Supplementary Information

Appendix Table 1 Exclusion Criteria

ICD-9 or ICD-10	Code	Description
ICD-10	F31.*	bipolar disorder
	F20.*	schizophrenia
	F23.*	brief psychotic disorder
	F25.*	schizoaffective disorder
	F32.3	mood [affective] disorders with psychotic symptoms
	F33.3	mood [affective] disorders with psychotic symptoms
	F10.15*	schizophrenic reaction in alcoholism
	F10.25*	schizophrenic reaction in alcoholism
	F10.95*	schizophrenic reaction in alcoholism
	F06.2	schizophrenic reaction in brain disease
	F11-F19	psychoactive drug use
	F21	schizotypal disorder
	F03.*	dementia
	F42.*	OCD
	R46.81	Obsessive-compulsive behavior
	F60.5	Obsessive-compulsive personality disorder
	F43.1	PTSD
	F84.*	autistic disorder
	F30	Manic episode
	F30.1	Manic episode without psychotic symptoms
	F30.10	Manic episode without psychotic symptoms, unspecified
	F30.11	Manic episode without psychotic symptoms, mild
	F30.12	Manic episode without psychotic symptoms, moderate
	F30.13	Manic episode, severe, without psychotic symptoms
	F30.2	Manic episode, severe with psychotic symptoms
	F30.3	Manic episode in partial remission
	F30.4	Manic episode in full remission
	F30.8	Other manic episodes
	F30.9	Manic episode, unspecified
	F22*	Delusional disorders
	F24*	Shared psychotic disorder
	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

Appendix Table 1 (Continued)

ICD-9 or ICD-10	Code	Description
ICD-9	296.0	Manic disorder, single episode
	296.00	Bipolar I disorder, single manic episode
	296.01	Manic disorder, single episode, mild degree
	296.02	Manic disorder, single episode, moderate degree
	296.03	Severe bipolar I disorder, single manic episode without psychotic features
	296.04	Severe bipolar I disorder, single manic episode with psychotic features
	296.05	Manic disorder, single episode, in partial or unspecified remission
	296.06	Bipolar I disorder, single manic episode, in remission
	296.1	Manic disorder, recurrent episode
	296.10	Manic disorder, recurrent episode, unspecified degree
	296.11	Manic disorder, recurrent episode, mild degree
	296.12	Manic disorder, recurrent episode, moderate degree
	296.13	Manic disorder, recurrent episode, severe degree, without mention of psychotic behavior
	296.14	Manic disorder, recurrent episode, severe degree, specified as with psychotic behavior
	296.15	Manic disorder, recurrent episode, in partial or unspecified remission
	296.16	Manic disorder, recurrent episode, in full remission
	296.4	Bipolar affective disorder, manic
	296.40	Bipolar I disorder, most recent episode manic
	296.41	Mild bipolar I disorder, most recent episode manic
	296.42	Moderate bipolar I disorder, most recent episode manic
	296.43	Severe manic bipolar I disorder without psychotic features
	296.44	Bipolar affective disorder, manic, severe degree, specified as with psychotic behavior
	296.45	Bipolar affective disorder, manic, in partial or unspecified remission
	296.46	Bipolar affective disorder, manic, in full remission
	296.5	Bipolar affective disorder, depressed
	296.50	Bipolar affective disorder, depressed, unspecified degree
	296.51	Bipolar affective disorder, depressed, mild degree
	296.52	Bipolar affective disorder, depressed, moderate degree
	296.53	Bipolar affective disorder, depressed, severe degree, without mention of psychotic behavior
	296.54	Bipolar affective disorder, depressed, severe degree, specified as with psychotic behavior
	296.55	Bipolar affective disorder, depressed, in partial or unspecified remission
	296.56	Bipolar affective disorder, depressed, in full remission

Appendix Table 1 (Continued)

ICD-9 or ICD-10	Code	Description
ICD-9	296.6	Bipolar affective disorder, mixed
	296.60	Bipolar affective disorder, mixed, unspecified degree
	296.61	Bipolar affective disorder, mixed, mild degree
	296.62	Bipolar affective disorder, mixed, moderate degree
	296.63	Bipolar affective disorder, mixed, severe degree, without mention of psychotic behavior
	296.64	Bipolar affective disorder, mixed, severe degree, specified as with psychotic behavior
	296.65	Bipolar affective disorder, mixed, in partial or unspecified remission
	296.66	Bipolar affective disorder, mixed, in full remission
	296.7	Bipolar affective disorder, unspecified
	296.8	Manic-depressive psychosis, other and unspecified
	296.80	Manic-depressive psychosis, unspecified
	296.81	Atypical manic disorder
	296.89	Other manic-depressive psychosis

Bibliography

1 Belmaker RH, Agam G. Major Depressive Disorder. *N Engl J Med* 2008;**358**:55–68. doi:10.1056/NEJMra073096

2 Milaneschi Y, Lamers F, Peyrot WJ, *et al.* Polygenic dissection of major depression clinical heterogeneity. *Physiol Behav* 2016;**21**:516–22. doi:doi:10.1038/mp.2015.86.

3 Fava M, Kendler KS. Major Depressive Disorder. *Neuron* 2000;**28**:335–41. doi:10.1016/B978-012373947-6.00245-2

4 Singh AB, Bousman CA. Antidepressant pharmacogenetics. *Am J Psychiatry* 2017;**174**:417–8. doi:10.1176/appi.ajp.2017.17020173

5 Collins PY, Patel V, Joestl SS, *et al.* Grand challenges in global mental health. *Nature* 2011;**475**:27–30.

6 James SL, Abate D, Abate KH, *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 354 Diseases and Injuries for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 2018;**392**:1789–858. doi:10.1016/S0140-6736(18)32279-7

7 Kessler RC, Berglund P, Demler O, *et al.* The epidemiology of major depressive disorder: results from the national comorbidity survey replication (NCS-R). *JAMA* 2003;**289**:3095–105. doi:10.1097/00132578-200310000-00002

8 Demyttenaere K, Bruffaerts R, Posada-Villa J, *et al.* Prevalence, Severity, and Unmet Need for Treatment of Mental Disorders in the World Health Organization World Mental Health Surveys. *JAMA* 2004;**291**:2581–90.

9 Wong ML, Licinio J. Research and treatment approaches to depression. *Nat Rev Neurosci* 2001;**2**:343–51. doi:10.1038/35072566

10 Friedrich MJ. Depression is the leading cause of disability around the world. *J Am Med Assoc* 2017;**317**:1517. doi:doi:10.1001/jama.2018.19559

11 Geneva: World Health Organzation. Depression and Other Common Mental Disorders: Global Health Estimates. 2017.

12 Licinio J, Wong M-L. The pharmacogenomics of depression. *Pharmacogenomics J* 2001;**1**:175–7. doi:10.1038/sj.tpj.6500047

13 Sullivan PF, Daly MJ, Ripke S, *et al.* A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry* 2013;**18**:497–511. doi:10.1038/mp.2012.21

14 Shi J, Potash JB, Knowles JA, *et al.* Genome-wide association study of recurrent early-onset major depressive disorder. *Mol Psychiatry* 2011;**16**:193–201. doi:10.1038/mp.2009.124

15 Rietschel M, Mattheisen M, Frank J, *et al.* Genome-wide association-, replication-, and neuroimaging study implicates homer1 in the etiology of major depression. *Biol Psychiatry* 2010;**68**:578–85. doi:10.1016/j.biopsych.2010.05.038

16 Shyn SI, Shi J, Kraft JB, *et al.* Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. *Mol Psychiatry* 2011;**16**:202–15. doi:10.1038/mp.2009.125

17 Muglia P, Tozzi F, Galwey NW, *et al.* Genome-wide association study of recurrent major depressive disorder in two European case-control cohorts. *Mol Psychiatry* 2010;**15**:589–601. doi:10.1038/mp.2008.131

18 Sullivan PF, De Geus EJC, Willemsen G, *et al.* Genome-wide association for major depressive disorder: A possible role for the presynaptic protein piccolo. *Mol Psychiatry* 2009;**14**:359–75. doi:10.1038/mp.2008.125

19 Wray NR, Pergadia ML, Blackwood DHR, *et al.* Genome-wide association study of major depressive disorder: New results, meta-analysis, and lessons learned. *Mol Psychiatry* 2012;**17**:36–48. doi:10.1038/mp.2010.109

20 Kohli MA, Lucae S, Saemann PG, *et al.* The Neuronal Transporter Gene SLC6A15 Confers Risk to Major Depression. *Neuron* 2011;**70**:252–65. doi:10.1016/j.neuron.2011.04.005

21 Lewis CM, Ng MY, Butler AW, *et al.* Genome-wide association study of major recurrent depression in the U.K. population. *Am J Psychiatry* 2010;**167**:949–57. doi:10.1176/appi.ajp.2010.09091380

22 Wray NR, Ripke S, Mattheisen M, *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* 2018;**50**:1–41. doi:10.1038/s41588-018-0090-3.

23 Marshe VS, Maciukiewicz M, Rej S, *et al.* Norepinephrine transporter gene variants and remission from depression with venlafaxine treatment in older adults. *Am J Psychiatry* 2017;**174**:468–75. doi:10.1176/appi.ajp.2016.16050617

24 Lesch KP, Bengel D, Heils A, *et al.* Association of Anxiety-Related Traits with a Polymorphism in the Serotonin Transporter Gene Regulatory Region. *Science (80-)* 1996;**274**:1527–31. doi:10.1126/science.274.5292.1527

25 Hu XZ, Rush AJ, Charney D, *et al.* Association between a functional serotonin transporter promoter polymorphism and citalopram treatment in adult outpatients with major depression. *Arch Gen Psychiatry* 2007;**64**:783–92. doi:10.1001/archpsyc.64.7.783

26 Heils A, Teufel A, Petri S, *et al.* Allelic Variation of Human Serotonin Transporter Gene Expression. *J Neurochem* 1996;**66**:2621–4. doi:10.1046/j.1471-4159.1996.66062621.x

27 Caspi A, Sugden K, Moffitt TE, *et al.* Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science (80-)* 2003;**301**:386–9. doi:10.1126/science.1083968

28 Casacalenda N, Perry JC, Looper K. Remission in major depressive disorder: A comparison of pharmacotherapy, psychotherapy, and control conditions. *Am J Psychiatry* 2002;**159**:1354–60. doi:10.1176/appi.ajp.159.8.1354

29 Cleare A, Pariante CM, Young AH, *et al. Evidence-based guidelines for treating depressive disorders with antidepressants: A revision of the 2008 British Association for Psychopharmacology guidelines*. 2015. doi:10.1177/0269881115581093

30 Bauer M, Pfennig A, Severus E, *et al.* World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for Biological Treatment of Unipolar Depressive Disorders, Part 1: Update 2013 on the acute and continuation treatment of unipolar depressive disorders. *World J Biol Psychiatry* 2013;**14**:334–85. doi:10.3109/15622975.2013.804195

31 Rush AJ, Trivedi MH, Wisniewski SR, *et al.* Acute and Longer-Term Outcomes in Depressed Outpatients Requiring One or Several Treatment Steps: A STAR*D Report. *Am J Psychiatry* 2006;**163**:1905–17.http://ajp.psychiatryonline.org.proxy.hsl.ucdenver.edu/doi/pdf/10.1176/ajp.2006.163.11.1905

32 Trivedi MH, Rush AJ, Wisniewski SR, *et al.* Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: Implications for clinical practice. *Am J Psychiatry* 2006;**163**:28–40. doi:10.1176/appi.ajp.163.1.28

33 Blier P. Optimal use of antidepressants: When to act? *J Psychiatry Neurosci* 2009;**34**:80.

34 Khushboo SB. Antidepressants: Mechanism of Action, Toxicity and Possible Amelioration. *J Appl Biotechnol Bioeng* 2017;**3**:437–48. doi:10.15406/jabb.2017.03.00082

35 Berlanga C, Flores-Ramos M. Different gender response to serotonergic and noradrenergic antidepressants. A comparative study of the efficacy of citalopram and reboxetine. *J Affect Disord* 2006;**95**:119–23. doi:10.1016/j.jad.2006.04.029

36 Kornstein SG, Schatzberg AF, Thase ME, *et al.* Gender differences in treatment response to sertraline versus imipramine in chronic depression. *Am J Psychiatry* 2000;**157**:1445–52. doi:10.1176/appi.ajp.157.9.1445

37 Sramek JJ, Murphy MF, Cutler NR. Sex differences in the psychopharmacological treatment of depression. *Dialogues Clin Neurosci* 2016;**18**:447–57. doi:10.1016/j.rinp.2017.12.053

38 Townsend MK, Clish CB, Kraft P, *et al.* Reproducibility of metabolomic profiles among men and women in 2 large cohort studies. *Clin Chem* 2013;**59**:1657–67. doi:10.1373/clinchem.2012.199133

39 Young EA. Sex Differences in Response to Citalopram : A STAR * D Report. *J Psychiatry Res* 2009;**43**:503–11. doi:10.1016/j.jpsychires.2008.07.002.Sex

40 Frances AJ, Pincus HA, First B, *et al.* DSM-IV. *JAMA* 1994;**272**:828–9.

41 Hamilton M. A Rating Scale for Depression. *J Neurol Neurosurg Psychiat* 1960;**23**:56–62. doi:10.1136/jnnp.23.1.56

42 Trajković G, Starčević V, Latas M, *et al.* Reliability of the Hamilton Rating Scale for Depression: A meta-analysis over a period of 49years. *Psychiatry Res* 2011;**189**:1–9. doi:10.1016/j.psychres.2010.12.007

43 Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: Validity of a Bried Depression Severity Measure. *JGIM* 2001;**16**:606–13.

44 Trivedi MH, Rush AJ, Ibrahim HM, *et al.* The Inventory of Depressive Symptomatology, clinician rating (IDS-C) and self-report (IDS-SR), and the Quick Inventory Depressive Symptomatology, clinician rating (QIDS-C) and self-report (QIDS-SR) in public sector patients with mood disorders: A psychome. *Psychol Med* 2004;**34**:73–82. doi:10.1017/S0033291703001107

45 Rush JA, Trivedi MH, Ibrahim HM, *et al.* The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-Report (QIDS-SR): A Psychometric Evaluation in Patients with Chronic Major Depression. *Depression* 2003;**54**:573–83. doi:10.1016/S0006-3223(03)01866-8

46 Keller MB, McCullough JP, Klein DN, *et al.* A Comparison of Nefazodone, the Cognitive Behavioral-Analysis System of Psychotherapy, and their Combination for the Treatment of Chronic Depression. *N Engl J Med* 2000;**342**:1462–70.

47 Keller MB, Gelenberg AJ, Hirschfeld RMA, *et al.* The treatment of chronic depression, Part 2: A double-blind, randomized trial of sertraline and imipramine. *J Clin Psychiatry* 1998;**59**:598–607. doi:10.4088/JCP.v59n1107

48 Frank E, Karp J, Rush A. Efficacy of treatments for major depression. *Psychopharmacol Bull* 1993;**29**:457–75.

49 Rush AJ, Fava M, Wisniewski SR, *et al.* Sequenced treatment alternatives to relieve depression (STAR*D): Rationale and design. *Control Clin Trials* 2004;**25**:119–42. doi:10.1016/S0197-2456(03)00112-0

50 Sinyor M, Schaffer A, Levitt A. The Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial: A review. *Can J Psychiatry* 2010;**55**:126–35. doi:10.1177/070674371005500303

51 Lekman M, Paddock S, McMahon FJ. Pharmacogenetics of major depression: Insights from level 1 of the sequenced treatment alternatives to relieve depression (STAR*D) trial. *Mol Diagnosis Ther* 2008;**12**:321–30. doi:10.1007/BF03256297

52 Paddock S, Laje G, Charney D, *et al.* Association of GRIK4 with outcome of antidepressant treatment in the STAR*D cohort. *Am J Psychiatry* 2007;**164**:1181–8. doi:10.1176/appi.ajp.2007.06111790

53 Lekman M, Laje G, Charney D, *et al.* The FKBP5-Gene in Depression and Treatment Response-an Association Study in the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) Cohort. *Biol Psychiatry* 2008;**63**:1103–10. doi:10.1016/j.biopsych.2007.10.026

54 PGRN. Pharmacogenomics Research Network. www.pgrn.org (accessed 15 Jun 2020).

55 Giacomini K, Brett C, Altman R, *et al.* The Pharmacogenetics Research Network: From SNP Discovery to Clinical Drug Response. 2007;**81**:328–45. doi:10.1038/sj.clpt.6100087.The

56 Hewett M, Oliver DE, Rubin DL, *et al.* PharmGKB: The pharmacogenetics knowledge base. *Nucleic Acids Res* 2002;**30**:163–5. doi:10.1093/nar/30.1.163

57 Relling M V., Klein TE. CPIC: Clinical pharmacogenetics implementation consortium of the pharmacogenomics research network. *Clin Pharmacol Ther* 2011;**89**:464–7. doi:10.1038/clpt.2010.279

58 CPIC. Clinical Pharmacogenetics Implementation Consortium.

59 Hicks JK, Bishop JR, Sangkuhl K, *et al.* Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for CYP2D6 and CYP2C19 genotypes and dosing of selective serotonin reuptake inhibitors. *Clin Pharmacol Ther* 2015;**98**:127–34. doi:10.1002/cpt.147

60 Hicks JK, Sangkuhl K, Swen JJ, *et al.* Clinical pharmacogenetics implementation consortium guideline (CPIC) for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants: 2016 update. *Clin Pharmacol Ther* 2017;**102**:37–44. doi:10.1002/cpt.597

61 FDA US. Table of Pharmacogenetic Associations. 2021.https://www.fda.gov/medical-devices/precision-medicine/table-pharmacogenetic-associations

62 FDA US. Table of Pharmacogenomic Biomarkers in Drug Labeling. 2021.https://www.fda.gov/drugs/science-and-research-drugs/table-pharmacogenomic-biomarkers-drug-labeling

63 Bell GC, Crews KR, Wilkinson MR, *et al.* Development and use of active clinical decision support for preemptive pharmacogenomics. *J Am Med Informatics Assoc* 2014;**21**:93–9. doi:10.1136/amiajnl-2013-001993

64 Relling M V., Evans WE. Pharmacogenomics in the clinic. *Nature* 2015;**536**:343–50. doi:10.1016/j.physbeh.2017.03.040

65 Ramsey LB, Ong HH, Schildcrout JS, *et al.* Prescribing Prevalence of Medications With Potential Genotype-Guided Dosing in Pediatric Patients. *JAMA Netw open*

2020;**3**:e2029411. doi:10.1001/jamanetworkopen.2020.29411

66 Samwald M, Xu H, Blagec K, *et al.* Incidence of exposure of patients in the United States to multiple drugs for which pharmacogenomic guidelines are available. *PLoS One* 2016;**11**:1–17. doi:10.1371/journal.pone.0164972

67 Hicks JK, El Rouby N, Ong HH, *et al.* Opportunity for Genotype-Guided Prescribing Among Adult Patients in 11 US Health Systems. *Clin Pharmacol Ther* Published Online First: 2021. doi:10.1002/cpt.2161

68 Tansey KE, Guipponi M, Hu X, *et al.* Contribution of common genetic variants to antidepressant response. *Biol Psychiatry* 2013;**73**:679–82. doi:10.1016/j.biopsych.2012.10.030

69 Sullivan PF, Agrawal A, Bulik CM, *et al.* Psychiatric genomics: An update and an Agenda. *Am J Psychiatry* 2018;**175**:15–27. doi:10.1176/appi.ajp.2017.17030283

70 Dolin RH, Boxwala A, Shalaby J. A Pharmacogenomics Clinical Decision Support Service Based on FHIR and CDS Hooks. *Methods Inf Med* 2018;**57**:E115–23. doi:10.1055/s-0038-1676466

71 Rasmussen L V., Smith ME, Almaraz F, *et al.* An ancillary genomics system to support the return of pharmacogenomic results. *J Am Med Informatics Assoc* 2019;**26**:306–10. doi:10.1093/jamia/ocy187

72 Adams SM, Anderson KB, Coons JC, *et al.* Advancing pharmacogenomics education in the core pharmd curriculum through student personal genomic testing. *Am J Pharm Educ* 2016;**80**. doi:10.5688/ajpe8013

73 Klein ME, Parvez MM, Shin JG. Clinical Implementation of Pharmacogenomics for Personalized Precision Medicine: Barriers and Solutions. *J Pharm Sci* 2017;**106**:2368–79. doi:10.1016/j.xphs.2017.04.051

74 Manolio TA, Chisholm RL, Ozenberger B, *et al.* Implementing genomic medicine in the clinic: The future is here. *Genet Med* 2013;**15**:258–67. doi:10.1038/gim.2012.157

75 Rasmussen-Torvik LJ, Stallings SC, Gordon AS, *et al.* Design and Anticipated Outcomes of the eMERGE-PGx Project: A Multi-Center Pilot for Pre-Emptive Pharmacogenomics in Electronic Health Record Systems. *Clin Pharmacol Ther* 2014;**96**:482–9. doi:10.1016/j.physbeh.2017.03.040

76 Cavallari LH, Beitelshees AL, Blake K V., *et al.* The IGNITE Pharmacogenetics Working Group: An Opportunity for Building Evidence with Pharmacogenetic Implementation in a Real-World Setting. *Clin Transl Sci* 2017;**10**:143–6. doi:10.1111/cts.12456

77 Weitzel KW, Elsey AR, Langaee TY, *et al.* Clinical Pharmacogenetics Implementation: Approaches, Success, and Challenges. *Am J Med Genet C Semin Med Genet* 2014;**0**:56–67. doi:10.1002/ajmg.c.31390.Clinical

78 Weitzel KW, Alexander M, Bernhardt BA, *et al.* The IGNITE network: A model for genomic medicine implementation and research. *BMC Med Genomics* 2016;**9**:1–13. doi:10.1186/s12920-015-0162-5

79 Hall-Flavin DK, Winner JG, Allen JD, *et al.* Using a pharmacogenomic algorithm to guide the treatment of depression. *Transl Psychiatry* 2012;**2**. doi:10.1038/tp.2012.99

80 Bradley P, Shiekh M, Mehra V, *et al.* Improved efficacy with targeted pharmacogenetic-guided treatment of patients with depression and anxiety: A randomized clinical trial demonstrating clinical utility. *J Psychiatr Res* 2018;**96**:100–7. doi:10.1016/j.jpsychires.2017.09.024

81 Bousman CA, Arandjelovic K, Mancuso SG, *et al.* Pharmacogenetic tests and depressive symptom remission: A meta-analysis of randomized controlled trials. *Pharmacogenomics* 2019;**20**:37–47. doi:10.2217/pgs-2018-0142

82 Rosenblat JD, Lee Y, McIntyre RS. Does Pharmacogenomic Testing Improve Clinical Outcomes for Major Depressive Disorder? A Systematic Review of Clinical Trials and Cost-Effectiveness Studies. *J Clin Psychiatry* 2017;**78**:720—729. doi:10.4088/jcp.15r10583

83 Chen S, Chou WH, Blouin RA, *et al.* The cytochrome P450 2D6 (CYP2D6) enzyme polymorphism: Screening costs and influence on clinical outcomes in psychiatry. *Clin Pharmacol Ther* 1996;**60**:522–34. doi:10.1016/S0009-9236(96)90148-4

84 Chou WH, Yan F-X, De Leon J, *et al.* Extension of a pilot study: Impact from the cytochrome P450 2D6 polymorphism on outcome and costs associated with severe mental illness. *J Clin Psychopharmacol* 2000;**20**:246–51. doi:10.1097/00004714-200004000-00019 LK - http://huji-primo.hosted.exlibrisgroup.com/openurl/972HUJI/972HUJI_SP?sid=EMBASE&sid=EMBASE&issn=02710749&id=doi:10.1097%2F00004714-200004000-00019&atitle=Extension+of+a+pilot+study%3A+Impact+from+the+cytochrome+P450+2D6+polymorphism+on+outcome+and+costs+associated+with+severe+mental+illness&stitle=J.+Clin.+Psychopharmacol.&title=Journal+of+Clinical+Psychopharmacology&volume=20&issue=2&spage=246&epage=251&aulast=Chou&aufirst=Wen+Hwei&auinit=W.H.&aufull=Chou

85 Keller MB, Ryan ND, Strober M, *et al.* Efficacy of paroxetine in the treatment of adolescent major depression: A randomized, controlled trial. *J Am Acad Child Adolesc Psychiatry* 2001;**40**:762–72. doi:10.1097/00004583-200107000-00010

86 Ebrahim S, Bance S, Athale A, *et al.* Meta-analyses with industry involvement are massively published and report no caveats for antidepressants. *J Clin Epidemiol* 2016;**70**:155–63. doi:10.1016/j.jclinepi.2015.08.021

87 Bartlett VL, Dhruva SS, Shah ND, *et al.* Feasibility of Using Real-World Data to Replicate Clinical Trial Evidence. *JAMA Netw open* 2019;**2**:e1912869. doi:10.1001/jamanetworkopen.2019.12869

88 Bielinski SJ, Olson JE, Pathak J, *et al.* Preemptive genotyping for personalized medicine: Design of the right drug, right dose, right timedusing genomic data to individualize treatment protocol. *Mayo Clin Proc* 2014;**89**:25–33. doi:10.1016/j.mayocp.2013.10.021

89 Patel HB, Gandhi S. A Review on Big Data Analytics in Healthcare using Machine Learning Approaches. *Proc 2nd Int Conf Trends Electron Informatics* 2018.

90 Chekroud AM, Lane CE, Ross DA. Computational Psychiatry: Embracing Uncertainty and Focusing on Individuals, Not Averages. *Biol Psychiatry* 2017;**82**:1–5. doi:10.1016/j.physbeh.2017.03.040

91 Athreya AP, Iyer R, Wang L, *et al.* Integration of machine learning and pharmacogenomic biomarkers for predicting response to antidepressant treatment: Can computational intelligence be used to augment clinical assessments? *Pharmacogenomics* 2019;**20**:983–8. doi:10.2217/pgs-2019-0119

92 Iniesta R, Stahl D, McGuffin P. Machine learning, statistical learning and the future of biological research in psychiatry. *Psychol Med* 2016;**46**:2455–65. doi:10.1017/S0033291716001367

93 Chekroud AM, Zotti RJ, Shehzad Z, *et al.* Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry* 2016;**3**:243–50. doi:10.1016/S2215-0366(15)00471-X

94 Kautzky A, Baldinger P, Souery D, *et al.* The combined effect of genetic polymorphisms and clinical parameters on treatment outcome in treatment-resistant depression. *Eur Neuropsychopharmacol* 2015;**25**:441–53. doi:10.1016/j.euroneuro.2015.01.001

95 Athreya AP, Iyer RK, Neavin D, *et al.* Augmentation of Physician Assessments with Multi-Omics Enhances Predictability of Drug Response: A Case Study of Major Depressive Disorder. *IEEE Comput Intell Mag* 2018;**13**:20–31. doi:10.1016/j.physbeh.2017.03.040

96 Athreya AP, Neavin D, Carrillo-Roa T, *et al.* Pharmacogenomics-Driven Prediction of Antidepressant Treatment Outcomes: A Machine-Learning Approach With Multi-trial Replication. *Clin Pharmacol Ther* 2019;**106**:855–65. doi:10.1002/cpt.1482

97 Lin E, Kuo PH, Liu YL, *et al.* A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. *Front Psychiatry* 2018;**9**:1–10. doi:10.3389/fpsyt.2018.00290

98 Perlman K, Benrimoh D, Israel S, *et al.* A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. *J Affect Disord* 2019;**243**:503–15. doi:10.1016/j.jad.2018.09.067

99 Lee Y, Ragguett RM, Mansur RB, *et al.* Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *J Affect Disord* 2018;**241**:519–32. doi:10.1016/j.jad.2018.08.073

100 Spedicato GA. Discrete Time Markov Chains with R. *R J* 2017.

101 Wells BJ, Nowacki AS, Chagin K, *et al.* Strategies for Handling Missing Data in Electronic Health Record Derived Data. *eGEMs (Generating Evid Methods to Improv patient outcomes)* 2013;**1**:7. doi:10.13063/2327-9214.1035

102 Schafer JL. Multiple imputation: A primer. *Stat Methods Med Res* 1999;**8**:3–15. doi:10.1191/096228099671525676

103 Rubin DB. Multiple Imputation after 18+ Years. *J Am Stat Assoc* 1996;**91**:473–89. doi:10.1080/01621459.1996.10476908

104 Jakobsen JC, Gluud C, Wetterslev J, *et al.* When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts. *BMC Med Res Methodol* 2017;**17**:1–10. doi:10.1186/s12874-017-0442-1

105 White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat. Med. 2010;**30**:377–99. doi:10.1002/sim.4067

106 Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics* 2013;**14**. doi:10.1186/1471-2105-14-106

107 Sterne JAC, White IR, Carlin JB, *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;**338**. doi:10.1136/bmj.b2393

108 Dziura JD, Post LA, Zhao Q, *et al.* Strategies for dealing with missing data in clinical trials: From design to analysis. *Yale J Biol Med* 2013;**86**:343–58.

109 Madley-Dowd P, Hughes R, Tilling K, *et al.* The proportion of missing data should not be used to guide decisions on multiple imputation. *J Clin Epidemiol* 2019;**110**:63–73. doi:10.1016/j.jclinepi.2019.02.016

110 Rajkomar A, Hardt M, Howell MD, *et al.* Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;**169**:866–72. doi:10.7326/M18-1990

111 Chouldechova A, Roth A. The Frontiers of Fairness in Machine Learning. 2018;:1–13.http://arxiv.org/abs/1810.08810

112 Joseph G, Dohan D. Recruiting minorities where they receive care: Institutional barriers to cancer clinical trials recruitment in a safety-net hospital. *Contemp Clin Trials* 2009;**30**:552–9. doi:10.1016/j.cct.2009.06.009

113 Klein TE, Ritchie MD. PharmCAT: A Pharmacogenomics Clinical Annotation Tool. *Clin Pharmacol Ther* 2018;**104**:19–22. doi:10.1002/cpt.928

114 Gaedigk A, Ingelman-Sundberg M, Miller NA, *et al.* The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele

Nomenclature Database. *Clin Pharmacol Ther* 2018;**103**:399–401. doi:10.1002/cpt.910

115 Mammoliti A, Smirnov P, Safikhani Z, *et al.* Creating reproducible pharmacogenomic analysis pipelines. *bioRxiv* 2019;:1–7. doi:10.1101/614560

116 Jukić MM, Haslemo T, Molden E, *et al.* Impact of CYP2C19 genotype on escitalopram exposure and therapeutic failure: A retrospective study based on 2,087 patients. *Am J Psychiatry* 2018;**175**:463–70. doi:10.1176/appi.ajp.2017.17050550

117 Bråten LS, Haslemo T, Jukic MM, *et al.* Impact of CYP2C19 genotype on sertraline exposure in 1200 Scandinavian patients. *Neuropsychopharmacology* 2020;**45**:570–6. doi:10.1038/s41386-019-0554-x

118 Aldrich SL, Poweleit EA, Prows CA, *et al.* Influence of CYP2C19 metabolizer status on escitalopram/citalopram tolerability and response in youth with anxiety and depressive disorders. *Front Pharmacol* 2019;**10**:1–12. doi:10.3389/fphar.2019.00099

119 Banda JM, Seneviratne M, Hernandez-Boussard T, *et al.* Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annu Rev Biomed Data Sci* 2018;**1**:53–68. doi:10.1146/annurev-biodatasci-080917-013315

120 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *J Am Med Informatics Assoc* 2013;**20**:144–51. doi:10.1136/amiajnl-2011-000681

121 Hogan WR, Wagner MM. Accuracy of Data in Computer-based Patient Records. *J Am Med Informatics Assoc* 1997;**4**:342–55. doi:10.1136/jamia.1997.0040342

122 Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: Challenges, recent advances, and perspectives. *J Am Med Informatics Assoc* 2013;**20**. doi:10.1136/amiajnl-2013-002428

123 Newton KM, Peissig PL, Kho AN, *et al.* Validation of electronic medical record-based phenotyping algorithms: Results and lessons learned from the eMERGE network. *J Am Med Informatics Assoc* 2013;**20**:147–54. doi:10.1136/amiajnl-2012-000896

124 Gottesman O, Kuivaniemi H, Tromp G, *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: Past, present, and future. *Genet Med* 2013;**15**:761–71. doi:10.1038/gim.2013.72

125 Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Informatics Assoc* 2013;**20**:117–21. doi:10.1136/amiajnl-2012-001145

126 Wei WQ, Teixeira PL, Mo H, *et al.* Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Informatics Assoc* 2016;**23**:20–7. doi:10.1093/jamia/ocv130

127 Schmiedeskamp M, Harpe S, Polk R, *et al.* Use of International Classification of Diseases,

Ninth Revision Clinical Modification Codes and Medication Use Data to Identify Nosocomial Clostridium difficile Infection . *Infect Control Hosp Epidemiol* 2009;**30**:1070–6. doi:10.1086/606164

128 Ritchie MD, Denny JC, Crawford DC, *et al.* Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record. *Am J Hum Genet* 2010;**86**:560–72. doi:10.1016/j.ajhg.2010.03.003

129 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**:7329–36. doi:10.1073/pnas.1510502113

130 Kwon BC, Anand V, Severson KA, *et al.* DPVis: Visual Analytics with Hidden Markov Models for Disease Progression Pathways. *IEEE Trans Vis Comput Graph* 2020;:1–15. doi:10.1109/TVCG.2020.2985689

131 Sukkar R, Katz E, Zhang Y, *et al.* Disease progression modeling using Hidden Markov Models. *Proc Annu Int Conf IEEE Eng Med Biol Soc EMBS* 2012;:2845–8. doi:10.1109/EMBC.2012.6346556

132 Sampathkumar H, Chen XW, Luo B. Mining Adverse Drug Reactions from online healthcare forums using Hidden Markov Model. *BMC Med Inform Decis Mak* 2014;**14**:1–18. doi:10.1186/1472-6947-14-91

133 Chen X, Wang ZX, Pan XM. HIV-1 tropism prediction by the XGboost and HMM methods. *Sci Rep* 2019;**9**:1–8. doi:10.1038/s41598-019-46420-4

134 Liu YY, Li S, Li F, *et al.* Efficient learning of continuous-time hidden Markov models for disease progression. *Adv Neural Inf Process Syst* 2015;**2015-Janua**:3600–8.

135 Sun Z, Ghosh S, Li Y, *et al.* A probabilistic disease progression modeling approach and its application to integrated Huntington's disease observational data. *JAMIA Open* 2019;**2**:123–30. doi:10.1093/jamiaopen/ooy060

136 Elfeki A, Dekking M. A Markov Chain Model for Subsurface Characterization: Theory and Applications. *Math Geol* 2001;**33**:503–5. doi:10.1007/s11004-006-9037-9

137 Milea D, Verpillat P, Guelfucci F, *et al.* Prescription patterns of antidepressants: Findings from a US claims database. *Curr Med Res Opin* 2010;**26**:1343–53. doi:10.1185/03007991003772096

138 Olekhnovitch R, Hoertel N, Limosin F, *et al.* Using filled prescription sequences to rank antidepressants according to their acceptability in the general population: The Constances cohort. *J Psychiatr Res* 2020;**123**:72–80. doi:10.1016/j.jpsychires.2020.01.017

139 Sawada N, Uchida H, Suzuki T, *et al.* Persistence and compliance to antidepressant treatment in patients with depression: A chart review. *BMC Psychiatry* 2009;**9**:1–10. doi:10.1186/1471-244X-9-38

140 Cover T, Thomas J. *Elements of Information Theory*. 2nd ed. Wiley-Interscience, New Jersey 2006.

141 Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32. doi:10.1201/9780429469275-8

142 LeDell E, Gill N, Aiello S, *et al.* h2o: R Interface for the 'H2O' Scalable Machine Learning Platform. Published Online First: 2021.https://cran.r-project.org/package=h2o

143 Nelder JA, Wedderburn RWM. Generalized Linear Models. *J R Stat Soc* 1972;**135**:370–84.

144 Friedman J. Greedy Function Approximation : A Gradient Boosting Machine. *Ann Stat* 2001;**29**:1189–232.

145 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *KDD 2016* 2016;:785–94. doi:http://dx.doi.org/10.1145/2939672.2939785

146 Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**:436–44. doi:10.1038/nature14539

147 Von Korff M, Simon G. The relationship between pain and depression. *Br J Psychiatry* 1996;**168**:101–8. doi:10.1192/s0007125000298474

148 Fishbain DA, Cutler R, Rosomoff H, *et al.* Chronic Pain-Associated Depression: Antecedent or Consequence of Chronic Pain? A Review. *Clin J Pain* 1997;**13**:116–37.

149 Gottlieb SS, Khatta M, Friedmann E, *et al.* The influence of age, gender, and race on the prevalence of depression in heart failure patients. *J Am Coll Cardiol* 2004;**43**:1542–9. doi:10.1016/j.jacc.2003.10.064

150 Beauboeuf-Lafontant T. 'You have to show strength': An exploration of gender, race, and depression. *Gend Soc* 2007;**21**:28–51. doi:10.1177/0891243206294108

151 Roxburgh S. Untangling Inequalities: Gender, Race, and Socioeconomic Differences in Depression. *Sociol Forum* 2009;**24**. doi:10.1111/j.1573-7861.2009.01103.x

152 Assari S. Social determinants of depression: The intersections of race, gender, and socioeconomic status. *Brain Sci* 2017;**7**. doi:10.3390/brainsci7120156

153 Assari S, Lankarani MM. Association between stressful life events and depression; intersection of race and gender. *J Racial Ethn Heal Disparities* 2015;**3**:349–56. doi:10.1007/s40615-015-0160-5

154 Riolo SA, Nguyen TA, Greden JF, *et al.* Prevalence of depression by race/ethnicity: Findings from the national health and nutrition examination survey III. *Am J Public Health* 2005;**95**:998–1000. doi:10.2105/AJPH.2004.047225

155 Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak* 2019;**19**. doi:10.1186/s12911-

019-0874-0

156 Iniesta R, Hodgson K, Stahl D, *et al.* Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. *Sci Rep* 2018;**8**:1–9. doi:10.1038/s41598-018-23584-z

157 Chandrashekar G, Sahin F. A survey on feature selection methods. *Comput Electr Eng* 2014;**40**:16–28. doi:10.1016/j.compeleceng.2013.11.024

158 Team RC. R: A Language and Environment for Statistical Computing. Published Online First: 2021.https://www.r-project.org/

159 Grolemund G, Wickham H. Dates and Times Made Easy with {lubridate}. *J Stat Softw* 2011;**40**:1–25.

160 Wickham H, François R, Henry L, *et al.* dplyr: A Grammar of Data Manipulation. Published Online First: 2021.https://cran.r-project.org/package=dplyr

161 Neuwirth E. RColorBrewer: ColorBrewer Palettes. 2014.

162 Kolde R. pheatmap: Pretty Heatmaps. 2019.

163 Liaw A, Wiener M. Classification and Regression by randomForest. *R News* 2002;**2**:18–22.

164 Denaxas S, Direk K, Gonzalez-Izquierdo A, *et al.* Methods for enhancing the reproducibility of biomedical research findings using electronic health records. *BioData Min* 2017;**10**:1–19. doi:10.1186/s13040-017-0151-7

165 Rasmy L, Zheng WJ, Xu H, *et al.* A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. *J Biomed Inform* 2018;**84**:11–6. doi:10.1016/j.jbi.2018.06.011

166 Oh J, Makar M, Fusco C, *et al.* A Generalizable, Data-Driven Approach to Predict Daily Risk of Clostridium difficile Infection at Two Large Academic Health Centers. *nfect Control Hosp Epidemiol* 2018;**39**:425–33. doi:10.1017/ice.2018.16

167 Kingma DP, Welling M. Auto-Encoding Variational Bayes. *2nd Int Conf Learn Represent ICLR 2014 - Conf Track Proc* 2013;**1**:1–14.

168 Devlin J, Chang MW, Lee K, *et al.* BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol - Proc Conf* 2019;**1**:4171–86.

169 Li Y, Rao S, Solares JRA, *et al.* BEHRT: Transformer for Electronic Health Records. *Sci Rep* 2020;**10**:1–12. doi:10.1038/s41598-020-62922-y

170 Choi E, Bahadori MT, Kulas JA, *et al.* RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. *Adv Neural Inf Process Syst* 2016;:3512–20.