Computational psychiatry: machine learning for clinical decision support in the treatment of major depression

by

James Russell Andrew Benoit

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Psychiatry

University of Alberta

Abstract

The goal of this thesis is to contribute to the fields of data-driven medicine and computational psychiatry by attempting to demonstrate the viability of machine learning for use in psychiatry, specifically in predicting treatment outcomes for major depression. This is attempted in four ways:

1. Chapter 2 is original research (and an intended article) describing a way to use ML to produce a learned classifier that takes as input patient clinical features to predict symptom remission after eight weeks of a specific antidepressant therapy.
2. Chapter 3 is original research (and an intended article) describing a way to use automated machine learning software for predicting treatment response after eight weeks of antidepressant therapy.
3. Chapter 4 is a literature review updating the reader on progress in how machine learning has been applied in the fields of psychiatry and personalized medicine.
4. Chapter 5 is a viewpoint (and intended article) suggesting changes in psychiatric prescribing practice that will occur as a result of deploying machine learning tools.

Chapter 2 uses data from 11 of Pfizer's desvenlafaxine (DVS; trade name Pristiq) clinical trials to demonstrate the construction and use of a machine learned model for predicting treatment outcomes in depression after eight weeks of treatment. Results show that using pre-treatment baseline data comprising psychiatric scales, laboratory test data, demographic information, and medication-related data, is sufficient to produce a classifier capable of predicting symptom remission, defined as a Hamilton Depression Rating Scale (HAM-D) score of ≤ 7, with 69.0% accuracy, 6.9% above chance predictions ($p<0.05$).

Chapter 3 draws from the same dataset, using the automated machine learning software (RapidMiner) to train classifiers to predict treatment response, defined as a ≥ 50% reduction in symptoms, based on the HAM-D scale. Without including early response data, classifiers were only able to predict response at 58.90%; after including early response data, classifiers were able to predict response at 70.05% accuracy.

Chapter 4 is framed as a conceptual review of machine learning in personalized medicine and psychiatry, focusing on recent applications of machine learning software to psychiatric care challenges. It covers four domains: data access, movement away from traditional statistical models, knowledge translation (KT) & commercialization of machine learning technology, and futurism. Within these domains, the chapter examines the development of Electronic Medical Records (EMR's) as they relate to personalized medicine and the interaction of health data with developing technologies such as streaming data and data ownership, the interaction of health data and machine learning, the health implementation environment, and current mental health tools being deployed commercially.

Chapter 5 is a viewpoint focusing on how we anticipate machine learning will affect clinical prescribing practice. Currently, clinical trials focus on demonstrating population-level safety and efficacy of new antidepressant drugs, but do not account for variance between individual patients. Deployment of machine learning and learned tools in the clinic will give clinicians the ability to compare the probabilities of different antidepressants being effective while minimizing side-effect profiles, on a patient-by-patient basis. This chapter focuses on the possible downstream effects of clinical machine learning tool deployment at different levels of the healthcare environment.

Among these four chapters, the thesis attempts to demonstrate the viability of using machine learning for prediction of psychiatric treatment outcomes, and to articulate how the field of data-driven medicine is advancing quickly toward widespread use. This work has relevance for understanding ways in which machine learning, clinical practice, and future drug development in a transition to a future that will be characterized by a more data-driven, outcome-focused environment for individual patients.

Preface

This thesis is an original work by James Benoit. The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board, Project Name “Machine Learning Prediction of Response to Treatment in Major Depressive Disorder”, No. Pro00064974, April 29, 2016.

Dedication

To my family, for their encouragement and unwavering support during the writing of this thesis.

Acknowledgments

I would like to acknowledge Pfizer Canada Inc. for their generous provision of the clinical trial data for this thesis. Thanks to Alberta Innovates, Alberta Innovates: Technology Futures, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Department of Psychiatry at the University of Alberta, and the University of Alberta, for their generosity in funding this work through scholarship support. My team of co-supervisors comprised Drs. Dursun, Greenshaw, and Greiner. I would like to thank Dr. Russ Greiner for his time and patience in explaining many of the finer points of machine learning, Dr. Serdar Dursun for contributing his clinical knowledge to my understanding of depression, and Dr. Andy Greenshaw for sharing his broad knowledge of data-driven mental health, and committee member Dr. Matt Brown for his assistance in learning the Python programming language.

I would also like to thank the fine programming community at Stack Overflow for answering many of the technical questions that arose during software production.

Table of Contents

Abstract ii
Preface v
Dedication vi
Acknowledgments vii
Table of Contents viii
List of Tables x
List of Figures xi
Chapter 1. Introduction 1
1.1 Diagnosis of Depression 1
1.2 Machine Learning 9
1.3 Machine Learning Regulation 12
1.4 Thesis Statements 14
1.5 References 17
Chapter 2. Using machine learning to predict remission in patients with major depressive disorder treated with desvenlafaxine. 22
Abstract 23
2.1 Background 25
2.2 Methods 28
2.3 Results 33
2.4 Discussion 35
2.5 Funding 41
2.6 References 42
2.7 Supplementary Materials 48
Connection Between Chapters 2-3 56
Chapter 3. Using automated machine learning to predict response in major depressive disorder patients treated with desvenlafaxine. 57
Abstract 58
3.1 Background 60
3.2 Methods 64
3.3 Results 71
3.4 Discussion 76

3.5 References 80
3.6 Supplementary Materials 84
Connection Between Chapters 3-4 85
Chapter 4. A conceptual review of machine learning in personalized medicine with a focus on psychiatry 86
4.1 Introduction 87
4.2 Data Access 91
4.3 Movement away from traditional statistical models 100
4.4 KT & Private sector engagement & commercialization of software 105
4.5 Futurism & implications for ML in mental health 108
4.6 A note on search parameters 109
4.7 References 110
Connection Between Chapters 4-5 122
Chapter 5. From Efficacy to Accuracy: How Machine Learning Will Change Prescribing Practice in Depression 123
5.1 References 129
Chapter 6. Conclusion 131
6.1 References 135
Bibliography 137
Appendices 161
Appendix 1. Code for Chapter 2 Concure Classifier 161

List of Tables

Table 2.1. Clinical trial characteristics 29
Table 2.2. Dataset statistics: mean demographic information and HAM-D scores for training and holdout sets. 30
Table 2.3. List of 92 features included in the training dataset 48
Table 2.4. Accuracy comparison of classifiers. 55
Table 3.1. Dataset demographic information and HAM-D17 mean score averages. 65
Table 3.2. List of 109 features included in the training dataset 67
Table 3.3. Most accurate classifier performance measures 71
Table 3.4. GLM trained classifier features and coefficients 72
Table 3.5. Lower envelope composition and boundaries from Figure 3.3 74
Table 3.6. Desvenlafaxine clinical trial datasets 84

List of Figures

Figure 1.1. Reproduction of RDoC Matrix for Negative Valence Systems 7
Figure 2.1. Distinguishing between the learning process (here using the “Learner (Concure)”; top to bottom) and the performance process using the classifier produced by the Learner (here, “Classifier ($C_{Concure}$)”; left to right). 32
Figure 2.2. Holdout data bootstrap, mean accuracy= 69.0%, chance accuracy= 62.1%, 10000 samples, n= 377/run, p= 0.0025 35
Figure 2.3. Data splitting process for $C_{Concure}$ 52
Figure 3.1. Distinguishing between the learning process (top to bottom) and the performance process using the classifier produced by the Learner (left to right). 62
Figure 3.2 RapidMiner's Auto Model process 68
Figure 3.3. Cost Curve comparison of classifier performance 73
Figure 3.4. Significance testing for trained GLM classifier against trained FLM classifier 74
Figure 3.5. Significance testing for trained GLM classifier against trained deep learning classifier 75
Figure 4.1. Reproduction of Magic Quadrant for Data Science and Machine Learning Platforms 92
Figure 5.1. Use case example for classification accuracy-based prescription system 127

Chapter 1. Introduction

In diagnosis and treatment of major depressive disorder (MDD; also referred to as depression), there is no simple diagnostic blood test, scan, or questionnaire, or measure consistently and accurately applied to guide treatment choice. The lack of tests able to consistently predict treatment outcomes results in attempts at treating the patient's disorder that rely on incomplete information, informed primarily by physician experience, clinical interviews, and further interactions with the patient. There is a significant downside to continuing to use this method: 40-60% of psychiatric treatment attempts fail (Masand, 2003). This persists outside the normal clinical environment as well: individual antidepressants in clinical trials are only 63% effective (Masand, 2003; Gartlehner et al., 2011). When treatment fails, patients often discontinue treatment or lose faith in the ability of their physician to treat their condition (Bados, Balaguer & Saldaña, 2007; Olfson et al., 2009; Shamir, Szor & Melamed, 2010). This lack of reliable treatments has a significant impact. As a condition, depression is the most pervasive and contributes most to the global burden of disease: 350,000,000 cases exist worldwide, and it is responsible for 76.4 million years lost to disability (YLD), 10.3% of the total burden of disease (Smith, 2014).

1.1 Diagnosis of Depression

Historically, depression is recognized at least as far back as ancient Greece c. 460-357 BCE as a condition called melancholia (Kaplan, 2009), consisting of, "aversion to food, despondency, sleeplessness, irritability, and restlessness" (Hippocrates, 1923-1931). Hippocrates' theory centered on four "humours," defined as substances in the body regulating human behaviour: yellow bile, black bile, blood, and phlegm (Kaplan, 2009). Normative human behaviour was thought to stem from a balance between these substances in the body, while imbalances were

expressed as illnesses (Wikipedia contributors, 2019), with depression stemming from an excess of black bile attributed to a planetary influence, in this case Saturn (Kaplan, 2009).

Understanding of depression progressed through Jean-Philippe Esquirol's work in the mid-to-late 1800's, which connected the idea that mood disturbances underlie depression. This contrasted with previous views of depression as a primarily intellectual disorder: a form of insanity (i.e. disturbed thoughts and deranged reasoning) (Kaplan, 2009). This idea of distinguishing depression from other mental illnesses based on its mood-related symptoms was ushered into a more contemporary form by the psychiatrist Emil Kraepelin in the later 1800's, who introduced the idea of syndrome-based classification. This centred on grouping mental illnesses based on common patterns of symptoms, called syndromes (Lawlor, 2012). The idea of delineating diseases based on course and outcome was revolutionary, and would later be used as a central ideology around which modern psychiatric classifications systems were designed (Shorter, 2015).

In modern North American psychiatry, the diagnosis of depression is codified in three major systems, the most prominent of which is the American Psychiatric Association's DSM (Diagnostic and Statistical Manual of Mental Disorders; APA, 2013). These systems also include the ICD (International Classification of Disease; WHO, 1992), and RDoC (Research Domain Criteria; Insel, 2010). The first version of the DSM, the DSM-I, was released in 1952, as an adaptation of a 1943 Technical Bulletin outlining psychiatric nomenclature for the U.S. Army, "Medical 203," from the Office of the Surgeon General (Houts, 2000; APA, 2019). It classified mental disorders into two major categories, disorders of brain function and failure of the

individual to adjust to their circumstances; depression was classified in the DSM-I as a failure to adjust (Grob, 1991). As the DSM was developed and updated, the next major change in understanding depression was seen in the DSM-III. This involved defining depression with a more communicable and precise criteria, rather than a contextual approach that focused on the appropriateness of the patient's mood given their circumstances (e.g. if a family member died, low mood would be expected) (Horwitz, Wakefield & Lorenzo-Luaces, 2016). The American psychiatrist Robert Spitzer was responsible for chairing the committee in charge of DSM-III development. His view of the techne (structure) of psychiatric diagnosis leading to the DSM's approach can be summed up in his 2003 paper, as, "…having each clinician creatively adopt their own definitions is no solution and would inevitably lead to a diagnostic Tower of Babel" (First & Spitzer, 2003). However, Spitzer's view of the telos (essential purpose) of using such an approach was to enable physicians to communicate more efficiently: simply stating a diagnosis of "depression" captures not only the likely features expressed by a patient, but also rules out disorders (e.g. bipolar disorder), the range of treatments to be considered, and the patient's likely future outcomes (First & Spitzer, 2003).

The APA task force in charge of addressing weaknesses in the DSM-III during DSM-IV development was led by the American psychiatrist Allen Frances. This task force focused on generating an empirical basis for justifying changes to the disorder classification scheme via evidence from literature review, as well as increasing harmonization with ICD-10 structure (APA, 2019). The most notable change after developing the DSM-IV was the presence of "clinical significance" as a diagnostic requirement: patients now had to experience significant distress or impairment in a major domain of their life (e.g. social, occupational) (Wikipedia contributors, 2019).

Moving to the most recent incarnation of the DSM, the DSM-5 (APA, 2013), a diagnosis of depression requires five of the following nine symptoms as assessed by a clinician, including at least one of two primary symptoms related to mood or anhedonia (as denoted by a "*" below), persisting for at least two weeks:

1. *Depressed mood
2. *Loss of interest or pleasure
3. More than 5% change in body weight in a month
4. Insomnia/hypersomnia
5. Observable psychomotor change (agitation or retardation)
6. Fatigue/loss of energy
7. Inappropriate feelings of guilt/worthlessness
8. Reduced ability to concentrate or make decisions
9. Thoughts of death/suicide, suicide attempt, or specific plan for suicide

In addition, these symptoms must cause distress or impairment to normal functioning, cannot be attributed to the patient being affected by a substance's physiological effects, cannot be attributed to another medical condition, are not better accounted for as being part of schizophrenia or psychosis (and variations thereof), and do not exist in conjunction with symptoms of mania or hypomania.

The DSM-5's classification of depression differs from the other major disorder classification system, the ICD-10 (International Classification of Disease, 10th edition) (WHO, 1992). This version of the ICD is updated regularly, with the last major update occurring in 2016 (WHO, 2016). A diagnosis of depression is made under Section V: Mental and behavioural disorders, and is less codified than the DSM-5 system. A description of depression is given, followed by

diagnostic windows and characteristics of mild, moderate, and severe depressive disorders. A "mild" depression diagnosis made with the ICD-10 requires two to three of the following symptoms, while a "moderate" diagnosis requires four, and "severe" requires "several", and severe depression is checked for psychotic symptoms (e.g. hallucinations, delusions):

1. Lowering of mood
2. Reduction of energy
3. Decrease in activity
4. Reduced capacity for enjoyment
5. Reduced capacity for interest
6. Reduced capacity for concentration
7. Marked tiredness after even minimum effort
8. Sleep is disturbed
9. Appetite is diminished
10. Self-esteem is reduced
11. Self-confidence is reduced
12. Ideas of guilt or worthlessness are present
13. Loss of libido

The ICD-10 divides depression into mild, moderate, and severe categories, with the number, severity, and effect of symptoms on day-to-day life contributing to the severity of diagnosis. In addition, the presence of somatic symptoms and suicidal thoughts are considered hallmarks for cases of severe depression. This diagnosis excludes cases of adjustment disorder (changes to mood resulting from significant life changes or stressors), recurrent depressive disorder (repeated episodes of depression), and when combined with symptoms of conduct disorder (characterized by dissocial/aggressive/defiant conduct).

Both the ICD-10 and DSM-5 discuss depression in the context of the patient feeling distress, displaying impaired day-to-day functioning, and a lack of symptoms of mania. However, the ICD-10 does not account for the effects of drug abuse on producing depressive symptoms, nor does it ask physicians to consider whether depressive symptoms would fit better in the context of a schizophrenia or psychosis diagnosis.

However, there has been increasing debate surrounding the ability of DSM-5 criteria to produce consistent diagnoses: the inter-rater reliability for a DSM-5 depression diagnosis is questionable, with a Cohen's Kappa score of 0.28 (Regier et al., 2013; Lieblich et al., 2015). Cohen's Kappa is a statistical measure used to determine the degree of consensus reached between two individuals making a qualitative assessment using categorical scale items. In concrete terms, a Kappa of 0.28 means that two physicians will agree on a patient being diagnosed with depression in 4-15% of cases. This is a significant decrease from previous DSM releases: Cohen's Kappas for depression diagnoses from the DSM III and IV, are 0.65 and 0.67 (Brown et al., 2001).

First and Spitzer (2003) recognized that the DSM was to be used primarily as a clinical communication tool, and that multiple diagnoses were not to be used for inferring an underlying etiology. However, the concerns surrounding multiple diagnoses given for a constellation of symptoms that may have, "one or two underlying processes that are being expressed in a complex way" (First & Spitzer, 2003), and lack neuroscience-based evidence contributing to disorder etiology, gave rise to another classification scheme: the Research Domain Criteria (RDoC), developed by then-director of the National Institute for Mental Health, Dr. Thomas

Insel (Insel et al., 2010; Insel 2013). RDoC is composed primarily of a matrix that divides disorders into systems:

1. Negative valence systems
2. Positive valence systems
3. Cognitive systems
4. Social processes
5. Arousal and regulatory systems
6. Sensorimotor systems

These are then further divided by construct (e.g. Negative valence systems includes acute threat, potential threat, sustained threat, loss, and frustrative nonreward). Each construct is then divided into eight elements, each element constituting a specific type of evidence that contributes to our understanding of the disease (i.e. genes, molecules, cells, circuits, physiology, behavior, self-report, and paradigms) (NIMH, 2019b). See Fig. 1.1 for a reproduced version of the RDoC negative valence systems matrix.

Negative Valence Systems

Construct/Subconstruct	Genes Notice	Molecules	Cells	Circuits	Physiology	Behavior	Self-Report	Paradigms
Acute Threat ("Fear")		Elements	Elements	Elements	Elements	Elements	Elements	Elements
Potential Threat ("Anxiety")		Elements	Elements	Elements	Elements		Elements	Elements
Sustained Threat		Elements	Elements	Elements	Elements	Elements	Elements	
Loss		Elements		Elements	Elements	Elements	Elements	Elements
Frustrative Nonreward		Elements		Elements		Elements	Elements	Elements

Figure 1.1. Reproduction of RDoC Matrix for Negative Valence Systems

Figure 1.1 is reproduced from https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/constructs/rdoc-matrix.shtml

It should be noted that RDoC was developed with the intent of being a framework for researchers and clinical scientists to use as a guide for recruiting research study participants, and is not

intended to replace the DSM or ICD classification systems; rather, it is eventually intended to, "…inform diagnostic approaches using new laboratory procedures, behavioral assessments, and novel instruments to provide enhanced treatment and prevention interventions" (NIMH, 2019). At the current time, "the framework is directed toward constructs most germane to mental disorders, and makes no claim to span the entire gamut of functional behavior" (NIMH, 2019). The NIMH's discussion of how RDoC fits with current definitions of disorders states that it focuses on developing knowledge of individual disorder mechanisms informed by behavioural neuroscience and genetics. Furthermore, it may provide novel definitions of disorders in the future, but in the short term is intended to improve information available regarding treatment choices (NIMH, 2019).

Interestingly, the RDoC's evidence-based approach is being reflected in recent initiatives such as the European College for Neuropsychopharmacology's Neuroscience-based Nomenclature (ECNP NbN) (ECNP, 2019). NbN was developed to address patient confusion surrounding the naming of neuroscience-based medications, but takes a more aggressive approach compared to RDoC: it is intended for clinical use while development is ongoing (nbn2r.org, 2019). This classification schema proposes a nomenclature that, "reflects the current knowledge and understanding about the targeted neurotransmitter/ molecule/system being modified and the mode/mechanism of action" (nbn2r.org, 2019).

Given the uncertainty surrounding depression diagnosis and treatment options, it is imperative that emerging approaches to improving treatment selection in depression provide clinicians with objective tools that focus on patient outcomes. Using techniques in a subfield of artificial

intelligence (AI) called machine learning (ML) may offer a novel approach to improving treatment outcomes.

1.2 Machine Learning

Machine learning is a branch of artificial intelligence research concerned with using computational models capable of learning patterns from data in order to make predictions (SAS, 2019). Its primary goal is to optimize predictions according to some objective. Machine learning is split into three major fields: supervised learning, unsupervised learning, and reinforcement learning; each of these fields has a different optimization objective. Supervised learning is focused on minimizing classification error, unsupervised learning on maximizing the expected probability of data belonging to one class (given any number of classes), and reinforcement learning on maximizing an expected reward (e.g. a video game score) (van de Meent, 2018). Machine learning is concerned with creating intelligent machines that react like (or better than) humans. To differentiate the two, an example of AI that is not machine learning is an expert system, which uses human-programmed rules to make decisions. Different again is data mining, which like machine learning is concerned with finding patterns in data, but generally does not focus on optimizing predictions or making decisions. Instead, it focuses on generating useful descriptions of the data.

To make the differences between these fields concrete, consider some patient data labeled with treatment responses to a medication. Data mining could be used to cluster patient cases together, and output similarities between patients. An AI-based expert system could use certain features from each patient to determine which treatment to give. A machine learning program could use

patient labels to learn a pattern in the data that predicts treatment response in novel, unlabeled patients.

There are several important distinctions delineating machine learning and the more traditional statistical approach to modeling of using association studies. Association studies involve construction of a model of how the world is thought to work (i.e. a hypothesis), collecting data to test that hypothesis through an experiment, and determining whether the results of that experiment support the hypothesis (i.e. significance testing). Correlational association studies determine the strength of association between two or more variables (e.g. how does treatment response vary with age), in order to validate or refute the proposed model. Categorical association studies build a probability distribution of discrete values (e.g. how does treatment response vary with sex). There are two primary weaknesses of these two approaches being applied to personalized medicine. First, association studies work with distributions of data, rather than individual cases. This means that data is viewed as a sample of the underlying ground truth distribution that occurs at the population-level. This limits how individual cases can be assessed: it minimizes the contribution of valid cases that do not fit the population distribution. Second, the outcomes do not support actionable, objective predictions: the model tested is produced based on previous work attempting to answer a similar question, and tested against the sample distribution of data collected. Individual cases assessed with the proposed model are assessed in the context of the assumed distribution underlying the sample data collected.

Machine learning classification takes a different approach. I will limit discussion here to supervised machine learning classification, which is most relevant to this thesis. Unlike an association study, machine learning is not concerned with whether two (or more) variables are

correlated, but rather with how well an algorithm can learn patterns in data that can be used to generate correct predictions on novel data. Algorithms are processes followed during problem-solving operations; here, they will be discussed in the context of computer-based algorithms. An algorithm that has finished learning is called a trained classifier or learned classifier. Each patient or participant to be included in assessing a learned classifier must have data for an outcome variable with a known value. The outcome variable can have two or more classes (e.g. 0 and 1; respond and non-respond; treat or do not treat). While the training step is occurring, machine learning algorithms use data with visible classes (i.e. labeled data) to assess each produced classifier, before adjusting the parameters and assessing it again. Training finishes when no further performance improvement is seen over successive parameter adjustments. Trained classifier performance is then assessed by making predictions on novel, unlabeled data. A prediction here means assigning a class label to a data instance (i.e. a patient or participant; also called a case). During performance evaluation, the trained classifier is blinded to class labels: the true class of the novel data is known to the researcher and will be used to assess classifier performance, but is not seen by the trained classifier.

A hypothesis for machine learning may include a specific performance target, or simply be "above chance." This means that the classifier performs significantly better than simply assigning a random class to each data instance. If predictions are at chance level (analogous to the model simply guessing each prediction), the learned model has been unable to find a pattern of predictions. If above chance, the model has learned a useful pattern in the data.

However, there are caveats to taking a machine learning approach over an association study approach. Results, especially with complex trained models, may not as interpretable as an association study, because as the structure of the learned model increases in complexity, the contribution of each feature is more difficult to discern, as is the nature of feature-feature interactions. In addition, the amount of cases required to make accurate predictions increases as the complexity of the predictions being made increases. In other words, predicting treatment response to a complex disorder such as depression will require more patient cases to learn than learning to determine whether a line drawn on a piece of paper is horizontal or vertical.

1.3 Machine Learning Regulation

The FDA is a regulatory body in the United States that is broadly responsible for protecting consumers by assessing the safety, efficacy, and security of foods, drugs, electronics, medical devices, cosmetics, and tobacco (FDA, 2019b). Within medical devices, machine learned models are classified as Software as a Medical Device (SaMD), and include software for the treatment, diagnosis, cure, mitigation, or prevention of disease. Moving machine learned models from a research environment to being deployed in a clinical environment (i.e. “bench to bedside”) requires FDA approval in the United States. Health Canada has a similar scope compared to the FDA (Health Canada, 2017). The FDA clearly defines what does not constitute a SaMD: tools for administrative support or lifestyle enhancement, tools acting as electronic medical records, tools for data manipulation/visualization, and tools for specific, limited cases of clinical decision support (FDA, 2019).

The IMDRF (International Medical Device Regulators Forum) has proposed a risk categorization framework for SaMD’s, where tools fall into three levels of information significance: (1) diagnosis/treatment tools, (2) clinical management drivers, and (3) clinical management

information support. Within each category, tools are divided by the state of condition to which they are applied: critical, serious, and non-serious (Medical Device Working Group, 2014). Based on the combination of information significance and seriousness of condition, tools are assigned an impact level from I-IV, which defines the risk associated with deploying the tool in a clinical environment.

In an attempt to better define which tools need regulation, the IMDRF's proposal has been integrated into the FDA's proposed AI/ML regulatory framework for SaMD's. The motivation behind the creation of the FDA's framework is that currently, only "locked" algorithms- those where a given input will produce a deterministic (as opposed to stochastic) output- are regulated (and therefore, able to be deployed). Specifically, regulations currently require resubmission each time a change to the algorithm is made, and do not cover adaptive algorithms: those in which a given input may produce different outputs based on changes in the behaviour of the algorithm after it learns from new data available (FDA, 2019).

With these changes, the FDA hopes to harmonize AI/ML development around four development principles (FDA, 2019):

1. Quality control & software best practices
2. Safety & efficacy testing of premarket software
3. Incorporation of risk management practices & performance monitoring
4. Creation of a transparent system focusing on real-world performance reporting

As a result, adaptive machine learning algorithms are poised to become viable, regulated tools that can be deployed in clinical settings. Some outcomes of this deployment are still difficult to anticipate. For example, the effect of digital mental health apps on a digital therapeutic alliance

(i.e. between the patient and the app) is unclear: one review of the literature suggested that a lack of standardized outcome measurement to evaluate this alliance prevented conclusions from being drawn (Henson et al., 2019). Another suggested that it is simply too early to see the effects of these algorithmic approaches: although artificial intelligence and machine learning innovations directed at reducing suicide are currently nascent, clinical impact will be seen within 2-5 years (Torous et al., 2018).

1.4 Thesis Statements

The goal of this thesis is to contribute to the fields of data-driven medicine and computational psychiatry by attempting to demonstrate the viability of machine learning for use in psychiatry, specifically in predicting treatment outcomes for major depression. This is attempted in four ways:

1. Chapter 2 is original research (and an intended article) describing a way to use machine learning to produce a learned classifier that takes as input patient clinical features to predict symptom remission after eight weeks of a specific antidepressant therapy.
2. Chapter 3 is original research (and an intended article) describing a way to use automated machine learning software for predicting treatment response after eight weeks of antidepressant therapy.
3. Chapter 4 is a literature review updating the reader on progress in how machine learning has been applied in the fields of psychiatry and personalized medicine.
4. Chapter 5 is a viewpoint (and intended article) suggesting changes in psychiatric prescribing practice that will occur as a result of deploying machine learning tools.

Chapter 2 uses data from 11 of Pfizer's desvenlafaxine (DVS; trade name Pristiq) clinical trials to demonstrate the construction and use of a machine learned model for predicting treatment

outcomes in depression after eight weeks of treatment. Results show that using pre-treatment baseline data comprising psychiatric scales, laboratory test data, demographic information, and medication-related data, is sufficient to produce a classifier capable of predicting symptom remission, defined as a Hamilton Depression Rating Scale (HAM-D) score of ≤ 7, with 69.0% accuracy, 6.9% above chance predictions ($p<0.05$).

Chapter 3 draws from the same dataset, using the automated machine learning software (RapidMiner) to train classifiers to predict treatment response, defined as a $\geq 50\%$ reduction in symptoms, based on the HAM-D scale. Without including early response data, classifiers were only able to predict response at 58.90%; after including early response data, classifiers were able to predict response at 70.05% accuracy.

Chapter 4 is framed as a conceptual review of machine learning in personalized medicine and psychiatry, focusing on recent applications of machine learning software to psychiatric care challenges. It covers four domains: data access, movement away from traditional statistical models, knowledge translation (KT) & commercialization of machine learning technology, and futurism. Within these domains, the chapter examines the development of Electronic Medical Records (EMR's) as they relate to personalized medicine and the interaction of health data with developing technologies such as streaming data and data ownership, the interaction of health data and machine learning, the health implementation environment, and current mental health tools being deployed commercially.

Chapter 5 is a viewpoint focusing on how we anticipate machine learning will affect clinical prescribing practice. Currently, clinical trials focus on demonstrating population-level safety and efficacy of new antidepressant drugs, but do not account for variance between individual

patients. Deployment of machine learning and learned tools in the clinic will give clinicians the ability to compare the probabilities of different antidepressants being effective while minimizing and predicting side-effect profiles, on a patient-by-patient basis. This chapter focuses on the possible downstream effects of clinical machine learning tool deployment at different levels of the healthcare environment.

Among these four chapters, the thesis attempts to demonstrate the viability of using machine learning for prediction of psychiatric treatment outcomes, and to articulate how the field of data-driven medicine is advancing quickly toward widespread use. This work has relevance for understanding ways in which machine learning, clinical practice, and future drug development in a transition to a future that will be characterized by a more data-driven, outcome-focused environment for individual patients.

1.5 References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Pub; 2013. 991 p.
2. American Psychiatric Organization (APA). DSM History [Internet]. [cited 2019 Jun 10]. Available from: https://www.psychiatry.org/psychiatrists/practice/dsm/history-of-the-dsm
3. Bados A, Balaguer G, Saldaña C. The efficacy of cognitive-behavioral therapy and the problem of drop-out. J Clin Psychol. 2007 Jun;63(6):585–92.
4. Brown TA, Di Nardo PA, Lehman CL, Campbell LA. Reliability of DSM-IV anxiety and mood disorders: implications for the classification of emotional disorders. J Abnorm Psychol. 2001 Feb;110(1):49–58.
5. Burton R. The Anatomy of Melancholy: What it Is, with All the Kinds, Causes, Symptoms, Prognostics and Several Cures of it. John C. Nimmo; 1886. 558 p.
6. First MB, Spitzer RL. The DSM: Not Perfect, but Better Than the Alternative. Psychiatric Times [Internet]. 2003 Apr 1 [cited 2019 Jun 12]; Available from: https://www.psychiatrictimes.com/dsm-not-perfect-better-alternative
7. Gartlehner G, Hansen RA, Morgan LC, Thaler K, Lux L, Van Noord M, et al. Comparative benefits and harms of second-generation antidepressants for treating major depressive disorder: an updated meta-analysis. Ann Intern Med. 2011 Dec 6;155(11):772–85.
8. Grob GN. Origins of DSM-I: a study in appearance and reality. Am J Psychiatry. 1991 Apr;148(4):421–31.
9. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960 Feb;23:56–62.
10. Health Canada. Health Portfolio - Canada.ca [Internet]. 2017 [cited 2019 Jun 14]. Available from: https://www.canada.ca/en/health-canada/corporate/health-portfolio.html

11. Henson P, Wisniewski H, Hollis C, Keshavan M, Torous J. Digital mental health apps and the therapeutic alliance: initial review. BJPsych open [Internet]. 2019;5(1). Available from: https://www.cambridge.org/core/journals/bjpsych-open/article/digital-mental-health-apps-and-the-therapeutic-alliance-initial-review/84D2BF70EEA1EAD7E681FF012651B55E
12. Hippocrates. Works of Hippocrates, Vol. I–IV. (Trans. W. H. S. Jones & E. T. Withington). Cambridge, MA: Harvard University Press; 1923-1931.
13. Horwitz AV, Wakefield JC, Lorenzo-Luaces L. History of Depression. In: The Oxford Handbook of Mood Disorders. 2016. p. 1–24.
14. Houts AC. Fifty years of psychiatric nomenclature: reflections on the 1943 War Department Technical Bulletin, Medical 203. J Clin Psychol. 2000 Jul;56(7):935–67.
15. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry. 2010 Jul;167(7):748–51.
16. Insel T. Post by Former NIMH Director Thomas Insel: Transforming Diagnosis [Internet]. National Institute of Mental Health. 2013 [cited 2019 Feb 10]. Available from: https://www.nimh.nih.gov/about/directors/thomas-insel/blog/2013/transforming-diagnosis.shtml/index.shtml
17. Kaplan HI. Kaplan & Sadock's Comprehensive Textbook of Psychiatry. Wolters Kluwer Health/Lippincott Williams & Wilkins; 2009. 4520 p.
18. Lawlor C. From Melancholia to Prozac: A History of Depression. OUP Oxford; 2012. 265 p.

19. Lieblich SM, Castle DJ, Pantelis C, Hopwood M, Young AH, Everall IP. High heterogeneity and low reliability in the diagnosis of major depression will impair the development of new drugs. BJPsych Open. 2015 Oct;1(2):e5–7.

20. Masand PS. Tolerability and adherence issues in antidepressant therapy. Clin Ther. 2003 Aug;25(8):2289–304.

21. Medical Device (SaMD) Working Group. "Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations. In International Medical Device Regulators Forum; 2014.

22. nbn2r.org - NBN New Knowledge, New Nomenclature [Internet]. [cited 2019 Jun 12]. Available from: http://nbn2r.org/

23. Neuroscience-based Nomenclature [Internet]. European College for Neuropsychopharmacology (ECNP). [cited 2019 Jun 12]. Available from: https://www.ecnp.eu/research-innovation/nomenclature.aspx

24. NIMH » Discussion [Internet]. National Institute of Mental Health. [cited 2019 Jun 12]. Available from: https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/discussion.shtml

25. NIMH » RDoC Matrix [Internet]. National Institute of Mental Health. [cited 2019 Jun 12]. Available from: https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/constructs/rdoc-matrix.shtml

26. Olfson M, Mojtabai R, Sampson NA, Hwang I, Druss B, Wang PS, et al. Dropout from outpatient mental health care in the United States. Psychiatr Serv. 2009 Jul;60(7):898–907.

27. Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, et al. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. Am J Psychiatry. 2013 Jan;170(1):59–70.

28. SAS. Machine Learning: What it is and why it matters [Internet]. SAS Analytics Insights. 2019 [cited 2019 Jun 13]. Available from: https://www.sas.com/en_ca/insights/analytics/machine-learning.html

29. Shamir D, Szor H, Melamed Y. Dropout, early termination and detachment from a public psychiatric clinic. Psychiatr Danub. 2010 Mar;22(1):46–50.

30. Shorter E. The history of nosology and the rise of the Diagnostic and Statistical Manual of Mental Disorders. Dialogues Clin Neurosci. 2015 Mar;17(1):59–67.

31. Smith K. Mental health: a world of depression. Nature. 2014 Nov 13;515(7526):181.

32. Torous J, Larsen ME, Depp C, Cosco TD, Barnett I, Nock MK, et al. Smartphones, Sensors, and Machine Learning to Advance Real-Time Prediction and Interventions for Suicide Prevention: a Review of Current Progress and Next Steps. Curr Psychiatry Rep. 2018 Jun 28;20(7):51.

33. U.S. Food and Drug Administration. About FDA [Internet]. U.S. Food and Drug Administration (FDA). 2019b [cited 2019 Jun 14]. Available from: http://www.fda.gov/about-fda

34. US FDA Artificial Intelligence and Machine Learning Discussion Paper; 2019. Available from: https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/SoftwareasaMedicalDevice/UCM635052.pdf

35. van de Meent J-W. Unsupervised Machine Learning and Data Mining [Internet]. DS 5230 / DS 4420 Class Notes; 2018; Northeastern University. Available from: https://course.ccs.neu.edu/ds5230f18/assets/slides/ds5230-f18-lecture-01.pdf

36. Wikipedia contributors. Diagnostic and Statistical Manual of Mental Disorders [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Jun 12]. Available from: https://en.wikipedia.org/w/index.php?title=Diagnostic_and_Statistical_Manual_of_Mental_Disorders&oldid=898942134

37. Wikipedia contributors. Humorism [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Jun 7]. Available from: https://en.wikipedia.org/w/index.php?title=Humorism&oldid=899951997

38. World Health Organization (WHO). ICD-10 Version:2016 [Internet]. 2016 [cited 2019 Jun 12]. Available from: https://icd.who.int/browse10/2016/en

39. World Health Organization. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. Geneva: World Health Organization; 1992.

Chapter 2. Using machine learning to predict remission in patients with major depressive disorder treated with desvenlafaxine.

James RA Benoit, MA[1*], Serdar M Dursun, MD[1], Russell Greiner, PhD[2], Bo Cao, PhD[1], Matthew RG Brown, PhD[1], Raymond W Lam, MD[3], Philip Cowen, MD[4], Andrew J Greenshaw, PhD[1]

[1] Department of Psychiatry, University of Alberta, 1E1 Walter Mackenzie Health Sciences Centre, 8440 112 St NW, Edmonton, Alberta, Canada, T6G 2B7

[2] Department of Computing Science, University of Alberta, 2-32 Athabasca Hall, Edmonton, Alberta, Canada, T6G 2E8

[3] Department of Psychiatry, University of British Columbia, Detwiller Pavilion, 2255 Wesbrook Mall, Vancouver, BC, Canada, V6T 2A1

[4] Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, OX3 7JX

[*] Corresponding Author

Abstract

Background

Major depressive disorder (MDD) is a common and burdensome condition that has low rates of treatment success. Although antidepressant medications are effective for MDD, remission rates are low and patients often require several medication switches to achieve remission. Hence, selecting an effective antidepressant is primarily determined by trial and error. Techniques using machine learning hold potential for predicting treatment success with a particular medication. This study uses baseline clinical data in creating machine learning models that learn to predict remission status after desvenlafaxine (DVS) treatment.

Methods

We applied machine learning algorithms to data from 3776 MDD patients in 11 phase-III/IV clinical trials, to produce a model predicting symptom remission, defined as an 8-week Hamilton Depression Rating Scale (HAM-D) score of 7. We trained the model on a randomly selected 90% of the data (n=3399), then evaluated that learned model on a holdout set (n=377).

Outcomes

Our resulting classifier, a trained linear support vector machine (SVM), had a holdout set accuracy of 69.0%, significantly greater than the probability of classifying a patient correctly by chance. We demonstrate that this learning process is stable by repeatedly sampling part of the training dataset and running the learner on this sample, then evaluating the learned model on the non-sample instances of the training set; these runs had an average accuracy of 67.0% +/- 1.8%. Our model, based on 26 clinical features, proved sufficient to predict DVS remission significantly better than chance. This may allow more accurate use of DVS without waiting 8

weeks to determine treatment outcome, and may serve as a first step towards changing psychiatric care by incorporating clinical assistive technologies using machine learned models.

Funding

Data for this project were provided by Pfizer Inc. through a data sharing partnership with the University of Alberta.

2.1 Background

It is important for clinicians to identify the best treatment for patients with major depressive disorder (MDD). Unfortunately, when selecting a pharmacological treatment, there are currently no accepted, empirically-based clinical care strategies for determining which antidepressant is most likely to be efficacious and well-tolerated. Selecting an antidepressant generally relies on clinical features and side effect profile (Kennedy et al., 2016). However, meta-analyses of clinical trials for newer antidepressants found 37% of patients do not achieve response (a relative reduction in symptoms) and 53% do not achieve remission (expressing less than an absolute threshold of symptoms) following 6-12 weeks of treatment (Gartlehner et al., 2011). These are troubling statistics, especially as early effective treatment of depression may improve functional recovery outcomes (Habert et al., 2016), and each treatment failure increases the chance of overall failure and increases treatment times (Kennedy et al., 2016). Unfortunately, there are currently no reliable, well-validated tests that identify the best treatment for each patient, as we cannot accurately predict a patient's individual response to any antidepressant treatment. Hence, prescribing an effective antidepressant remains a trial-and-error process.

Many clinicians now rely on the DSM-5 framework of symptom clusters as a primary information source for diagnosis (APA, 2013). However, the DSM-5 does not incorporate patient genetics, physiology, nor other domains of information that may contribute to improved diagnostic stratification and treatment recommendation, such as those found in other diagnostic approaches such as the Research Domain Criteria (RDoC) (Insel et al., 2010). In addition, test-retest reliability for many DSM-5 diagnoses is questionable: studies have found less than 25% agreement between interviewers for the diagnosis of MDD in DSM-5 field trials (Regier et al., 2013). While multi-domain diagnosis is still early in development and unlikely to replace the

DSM-5 in the near future, guidelines for personalized treatment plans to enhance treatment efficacy are beginning to emerge (Oluboka et al., 2018). In this study, the clinical trials used were conducted between 2003-2011, using the previous version of the DSM, the DSM-IV-TR (APA, 2000) for diagnosis, which showed better agreement between physicians between 35-63% for MDD (kappa of 0.67; Brown et al., 2001).

Precision medicine attempts to identify which specific patients will respond to each specified treatment using models that can incorporate all available patient information. This approach uses outcomes, rather than symptom clusters, to divide patients into treatment groups, allowing for a data-driven approach. Machine learning, a subfield of artificial intelligence, includes techniques that lead to a precision medicine approach, as they are able to create accurate models of pharmacotherapy response, using potentially any type of patient information, including easily collected clinical measures (e.g. demographics, Hamilton Depression Rating Scale (HAM-D) items) (Chekroud et al., 2016; Hamilton, 1960). A major focus of machine learning in psychiatry has been producing models that diagnose mental health disorders, using neuroimaging data, including variants of Magnetic Resonance Imaging (MRI; e.g. the ADHD-200 competition) (Brown et al., 2012). Using machine learning tools to predict medication efficacy using patient information would move prescribing from inferential, trial-and-error practice to more precision medicine.

In this study, we applied machine learning to a large, global, multi-site dataset from eleven phase-III/IV clinical trials of the serotonin and norepinephrine reuptake inhibitor (SNRI), desvenlafaxine succinate (DVS). DVS is the primary active metabolite of the SNRI venlafaxine (thereby avoiding venlafaxine's interaction and metabolism by the liver enzyme CYP2D6),

acting as a reuptake inhibitor for both serotonin and norepinephrine with minimal effect on dopamine (Deecher et al., 2006).

The objective of this study was to develop a predictive model for treatment remission using baseline clinical information. This model's performance was evaluated on novel data, and the stability of its predictive accuracy confirmed on subsamples of the dataset.

2.1.1 Research in Context

2.1.1.1 Evidence before this study

Predicting symptom remission is important in MDD because of the high (>50%) number of patients who fail to remit following treatment. We searched PubMed from inception to Feb 15 2019, using the string (depression OR "major depressive disorder" OR MDD) AND ("machine learning" OR "treatment prediction" OR "response prediction"), with text available in any language. Of the 295 articles retrieved, we reviewed abstracts of the 78 where depress* AND predict* were in the title, and read full text articles based on abstract relevance.

We found that there are two groups of studies: those working with small, in-house datasets, versus others using large-scale databases such as the STAR*D; 11 had a sample size >500. The most-used scales in predictive outcome assessment were the HAM-D, Montgomery-Åsberg Depression Rating Scale (MADRS), and Quick Inventory of Depressive Symptomatology (QIDS) scales. We identified no other studies that synthesized multiple clinical trials into a single dataset for building predictive models of treatment response, nor built a model of treatment response prediction for DVS.

2.1.1.2 Added value of this study

First, we introduced a competing models approach to find the most effective of a group of machine learning algorithms at predicting patient response. Second, we expanded our scope beyond a large single-country trial (STAR*D) to using global data from multiple clinical trials, spanning 23 countries and 5 continents, with a large single-drug sample. We also expanded on previous methods of feature selection by applying a consistency-based feature selection method, which reduced the initial set of 92 clinical features down to 26 features, while demonstrating that the features picked were consistent across subsets of data.

2.1.1.3 Implications of all the available evidence

We show (1) that it is possible to synthesize multiple clinical trials into a large single dataset that can be used effectively for creating predictive models of MDD symptom remission; and (2) that the application of machine learning techniques to multimodal clinical trial data (psychiatric scales, lab tests, and demographic data) is beneficial for predicting symptom remission in MDD.

2.2 Methods

2.2.1 Datasets

The clinical trial data included in this study were drawn from 11 DVS clinical trials. We selected studies that were completed phase III/IV DVS trials with adult participants, and had a Hamilton Depression Rating Scale (HAM-D) outcome measure (Hamilton, 1960). Data were obtained through a data access agreement between Pfizer Inc. and the University of Alberta. This study was approved by the University of Alberta Research Ethics Board, study Pro00064974, and all patients involved gave written consent for their anonymized data to be used.

As shown in Table 2.1, our dataset combined data from 11 phase-III/IV DVS clinical trials carried out between 2003 and 2011, with a total enrollment of 7051 patients. After data cleaning and preprocessing, our dataset included 3776 patients: a training set of 3399 patients, and a holdout set of 377 patients (a randomly-selected 10% of the participant group was held aside from the machine learning process; see Table 2). The primary reason behind the reduction from 7051 patients to 3776 patients was a lack of week-8 HAM-D score, due to missing data or patient drop-out.

Table 2.1. **Clinical trial characteristics**

Dataset	Locations	DVS Remit % (included subjects)	Year
NCT01309542	Estonia, Finland, Former Serbia and Montenegro, France, Germany, Latvia, Lithuania, Poland, Slovakia, South Africa, United States	56.7	2003-2006
NCT00384033	United States	26.3	2006-2007
NCT00445679	China, India, Republic of Korea, Taiwan	45.9	2007-2009
NCT00406640	Argentina, Chile, Colombia, Mexico, Peru, United States	43.7	2006-2008
NCT00369343	United States	40.6	2006-2008
NCT00798707	Japan, United States	20.8	2008-2010
NCT00863798	United States	20.3	2009-2010
NCT01121484	United States	23.5	2010-2011
NCT00824291	United States, Canada	38.0	2009

NCT00300378	Croatia, Estonia, Finland, France, Latvia, Lithuania, Poland, Romania, Slovakia, South Africa	35.1	2006-2007
NCT00277823	United States	41.2	2006-2007

2.2.2 Inclusion criteria

Patient inclusion criteria were: a primary diagnosis of MDD, treated in a DVS monotherapy arm of a trial, and completion of a 17-item HAM-D assessment at both baseline and 8 weeks. Patients were excluded if they had comorbid psychiatric diagnoses.

Table 2.2. Dataset statistics: mean demographic information and HAM-D scores for training and holdout sets.

	Training	Holdout
n	**3399**	**377**
Age (years)	**44.0**	**43.6**
Sex (% Female)	**69.8**	**67.9**
Ethnicity (% White)	**65.1**	**65.0**
HAM-D Baseline	**21.3**	**21.3**
HAM-D Week 8	**10.9**	**11.3**
Remission Rate %	**37.9**	**37.9**

2.2.3 Outcome measures

We assessed treatment outcomes according to the clinician-reported 17-item HAM-D, obtained 8-weeks after the start of the trial, with the key outcome symptom of remission defined by a HAM-D score of 7 (Trivedi et al., 2006).

2.2.4 Features considered

Our training dataset D included 92 features, whose values were known for each patient at the start of the trial: psychiatric scale items (i.e. individual items from the Clinical Global Impressions Scale (CGI) (Guy, 1976), MADRS (Montgomery & Asberg, 1979), and HAM-D, demographic data (e.g. age, ethnicity), lab tests (e.g. free T4, white blood cell count), and adjunct medications including non-prescription drugs and supplements, summarized by a single feature indicating degree of polypharmacy at baseline: "How many different pills do you take each day?" We filled missing data points (e.g. a patient was missing the value for age), using mean imputation.

2.2.5 Predictive model

We applied our machine learning algorithm, which we call Concure (as it chose consistently picked features), to the labeled training dataset, D. This training dataset describes each patient using a set of clinical features, drawn from baseline measures taken when that patient entered the clinical trial. Each patient has a label of either "Remit" or "Non Remit", depending on whether that patient remitted at 8 weeks, indicated by a HAM-D score of 7. Using the training data, Concure returned a trained classifier, called $C_{Concure}(.)$, that predicts whether a novel patient, with his or her own values for these features, would experience symptom remission at 8 weeks. $C_{Concure}$(Patient A) = Remit means $C_{Concure}$ predicts patient A will remit, while $C_{Concure}$(Patient B) = Non-Remit means $C_{Concure}$ predicts patient B will not remit. Note that the Concure learner trains a classifier based on a set of labeled training data (here, D). That classifier then predicts a label for a novel patient; see Figure 1.

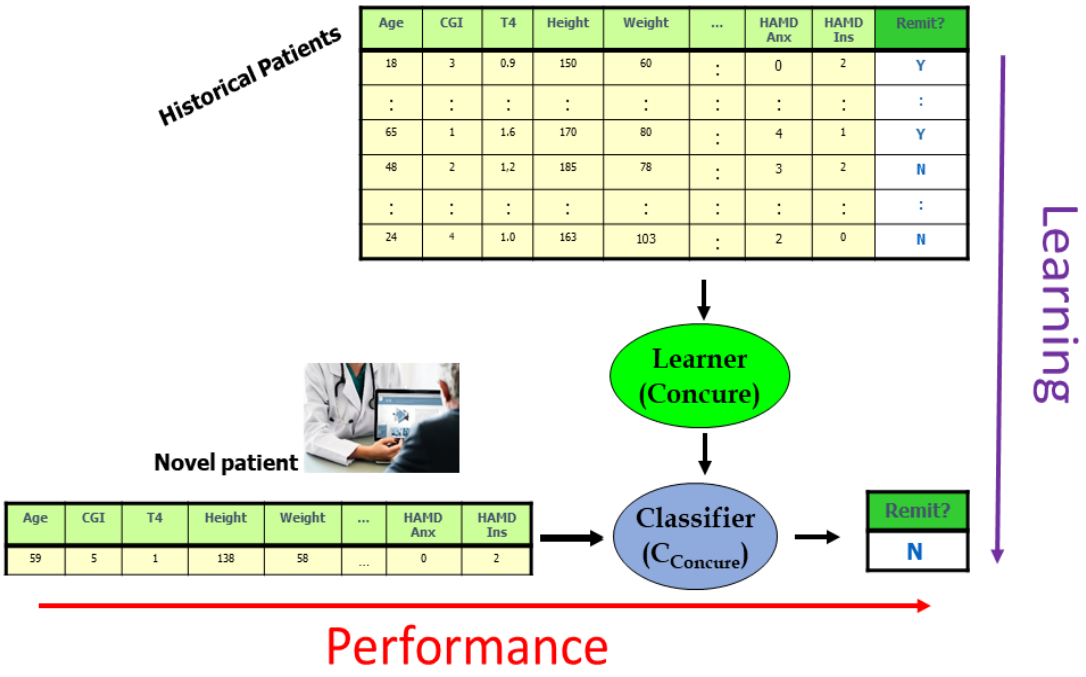

Figure 2.1. Distinguishing between the learning process (here using the "Learner (Concure)"; top to bottom) and the performance process using the classifier produced by the Learner (here, "Classifier ($C_{Concure}$)"; left to right).

The Concure learner involves 3 sequential steps: (1) Identify the subset of features f* that appear most informative for predicting remission vs non-remission; (2) identify the best "base learner," BL^* using these features (described below); and (3) run that BL^* on the dataset projected onto the feature subset f*. Concure initially selects features from each training fold's data using Lasso, then takes the intersection of these features to form a feature subset f*. Concure then considered 11 base learners (each is an algorithm that produces a classifier from a training dataset; see Appendix) and found that a linear support vector machine (SVM) classifier did best. It then ran a linear SVM learner on all the labeled training data, using the feature set f*, to produce a final trained classifier, $C_{Concure}$ (see section 2.7, Supplementary Materials, for details of this process).

We estimate $C_{Concure}$'s predictive accuracy in two ways. We chose "accuracy" (see section 2.7, Supplementary Materials, for equation) as our performance measure as it equally weights type I and II error. First, we use (external) cross-validation over the training data, D: Here, we run the entire Concure learning process (including the feature selection (Brown G et al., 2012), base learner selection, etc.) five times, each time on ⅘ of D, and evaluate that classifier on the remaining ⅕ of D (hence, evaluating it on the portion that it was not trained on). We report the average as an estimate. (Section 2.7, Supplementary Materials, provides a formal description of this process.)

Second, we applied the trained $C_{Concure}$ classifier to our patient holdout dataset to determine whether this model generalizes to novel patients from datasets that were entirely separate from the dataset used to train the classifier. This returned a single accuracy value. To assess whether that accuracy value is significantly different from chance, we used bootstrapping, based on 10,000 draws-with-replacement of size n=377 from the holdout set. The p-value for significance is determined by computing the percentage of sample means falling below the "chance probability" of correctly classifying a patient by guessing that all patients were non-remitters (here, 62.1%, corresponding to the majority class of patients, non-remitters).

2.3 Results

2.3.1 Feature selection

The feature subset f*was found to include 26 features (grouped by feature type):

1) Nine Countries of Origin (with each country considered as an individual binary feature): Argentina, Canada, China, Colombia, Croatia, Finland, Japan, Mexico, and USA.
2) One Ethnicity (American Indian/Alaska Native)
3) Eight HAM-D Scale Items:
    i) Anxiety/Somatic (anxiety concomitants, e.g. headaches, sweating)

ii) Feelings of Guilt (including rumination, delusions, hallucinations of guilt)
iii) Genital Symptoms
iv) Loss of Insight
v) Insomnia/Early (difficulty falling asleep)
vi) Somatic Symptoms/Gastrointestinal
vii) Somatic Symptoms/General (e.g. muscle ache, loss of energy, fatigability)
viii) Work & Activities (e.g. difficulty working or doing hobbies, being productive)

4) Three MADRS Scale Items:
    i) Apparent sadness
    ii) Pessimistic thoughts
    iii) Reported sadness
5) One measure of Polypharmacy (medication count including supplements, non-prescription drugs)
6) Four lab tests:
    i) Albumin
    ii) Creatinine
    iii) Potassium
    iv) Urine pH

2.3.2 Classifier selection

The classifier learned by the SVM base learner, C_{SVM}, was consistently the most accurate of the 11 trained classifiers tested, in the internal cross-validation folds. This classifier considers the 26 selected features in 26-dimensional space and generates a hyperplane that best separates the two classes.

2.3.3 Estimating the quality of Concure

As mentioned above, we evaluated our results in two ways. First, the 5-fold cross-validation accuracy (with respect to the training set) was 67.0% +/- 1.8% (SD). A two-tailed t-test shows this is significantly different from the 62.1% chance accuracy, p=0.0065.

2.3.4 Model validation, for $C_{Concure}$

Second, to explore the external generalizability of our learned 26-item $C_{Concure}$ model, we tested it on holdout data (n= 377, 37.9% remitters, 62.1% non-remitters). Its mean accuracy was 69.0%. We built an empirical distribution based on 10,000 bootstrap samples (SD = 2.4%), which was significantly different from chance accuracy at p = 0.0025; see Fig. 3.

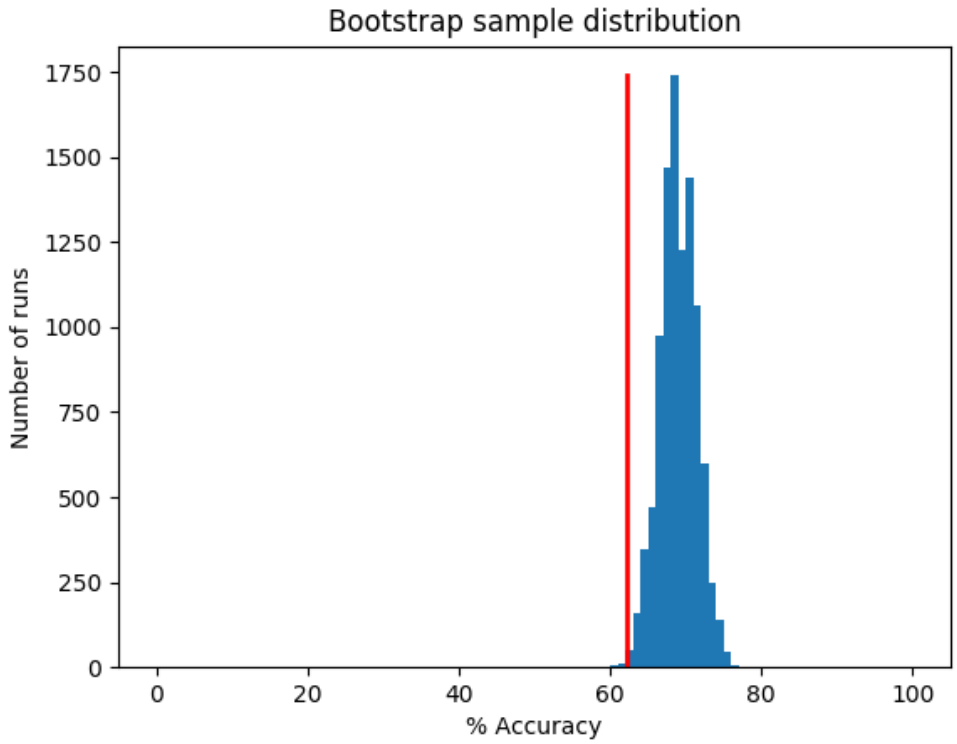

Figure 2.2. Holdout data bootstrap, mean accuracy= 69.0%, chance accuracy= 62.1%, 10000 samples, n= 377/run, p= 0.0025

2.4 Discussion

Our Concure learning algorithm produced a classifier capable of identifying, with better-than-chance performance, whether new patients diagnosed with MDD will experience symptom remission after 8 weeks of DVS monotherapy. This classifier demonstrates that a simple model, using 26 easily-obtained clinical features at baseline, can predict symptom remission at better

than chance levels, even when applied to a heterogeneous holdout dataset not used to train the classifier.

For comparison to other studies, the clinical features suggested in Chekroud et al.'s (2016) model include six HAM-D items: overall score, loss of insight, somatic energy (equivalent to somatic symptoms/general), somatic anxiety, delayed insomnia, and suicidality. Even though their model was predicting response to a different antidepressant, citalopram, our models shared three HAM-D features: loss of insight, somatic anxiety, and somatic energy. Interestingly if we split HAM-D items into four previously proposed symptom clusters based on principal component analysis (mood, sleep/psychic anxiety, weight/somatic anxiety, and insight/appetite) (Trivedi et al., 2005), both models contain HAM-D items from all four clusters. This may suggest that predicting treatment outcomes will be strongest in models that capture and consider multiple MDD subtypes.

Comparing our model to Iniesta et al.'s combined outcome prediction model for escitalopram and nortriptyline (Iniesta et al., 2016), our model shared MADRS apparent sadness and HAM-D work & activities. Apparent sadness relates to a core feature of MDD (mood) and is therefore an expected feature to be included in models of treatment outcome, and workplace functioning has previously been shown to be improved by both DVS and escitalopram (Lee et al., 2018).

Because our machine learning approach is agnostic in its consideration of features, it chose a disparate set of both expected and unexpected features. Of the 26 features selected, 11 were items from well-validated psychiatric scales: 8 based on HAM-D items, and 3 on MADRS items.

It also included nine countries of origin, one ethnicity feature, four lab tests, and one measure of polypharmacy.

Inclusion of the HAM-D "loss of insight" item is unexpected, as it is the least frequently occurring symptom of depression at baseline, and shows the least change of any item at treatment termination (McIntyre et al., 2002). However, this result is consistent with the findings of Chekroud et al. (Chekroud et al., 2016), suggesting that the machine learner finds value in including this feature across datasets and methodologies.

The polypharmacy feature was also used as a predictor of DVS efficacy, even though the simple nature of the item (number of pills taken daily by the patient) does not allow a fine-grained model for each adjunct medication. As there were 1507 different medications and other supplements listed that varied across patients, adding each one to the model would likely increase the risk of overfitting.

The lab findings were also unexpected features based on previous work. Albumin is typically not associated with remission from MDD: it is only mentioned in studies examining depression in patients with comorbid advanced kidney disease and cancer (Jhamb et al., 2018). Increased creatinine level has been associated with increased polypharmacy in older adults (Ersoy & Engin, 2018). As polypharmacy is also a feature in our machine learning prediction model, these may represent a latent predictive feature underlying both. Urine potassium has been weakly associated with fatigue and cortisol levels in subjects on a low-sodium-high-potassium diet (Torres, Nowson & Worsley, 2008), but there is currently no evidence that links it directly to remission from MDD. Similarly, the literature does not link urine pH with depression.

In addition, the specific features used by Concure are very convenient, as they are easily obtainable in low income jurisdictions and marginalized populations, with limited or no access to advanced medical technology such as MRI.

We had a choice of using MADRS, CGI, or HAM-D scales to assess patient outcome. We selected the HAM-D since it has been a "gold standard" for 40+ years of MDD research (Bagby et al., 2004), and is one of the three FDA-accepted endpoints for assessing antidepressant efficacy (CDER, 2018). We hope to expand this method in future studies in two ways (with the caveat that each would require a demonstration of construct validity). First, testing scale-based outcomes against patient self-assessments of remission would allow us to create a better proxy measure conducive to predictive modeling. Second, using as outcome a label that combines symptomatic and functional assessment (e.g. HAM-D and Sheehan Disability Scale) would allow a combined outcome that incorporates both functional and symptomatic remission, giving a more complete picture of whether a patient responded to treatment (Sheehan et al., 2017).

While the features selected have proven to be sufficient for significantly above-chance predictions, this analysis does not show them to be causally related to remission of depression (Pearl, 2009). While the literature has described associations between these features and depression treatment response, a different learning process (on this or a similar dataset) might select an entirely different set of features. That is not to say the features selected are irrelevant: given novel patient data, the analysis suggests that our trained classifier should accurately predict remission in 69.0% of patients taking DVS.

Other work in predicting MDD treatment outcomes has focused on applying a single learned algorithm to multiple cohorts. While useful for external validation of the model, those authors suggested that the models learned were specific to the mechanism underlying a particular treatment and might not generalize well across medications.

We had to contend with missing data, which we addressed in a very simple way: mean imputation. We tried an alternative strategy, median imputation, but found this preprocessing step did not lead to classifiers that could accurately classify new patients as remitters versus non-remitters. In some cases, we excluded features missing entirely from some datasets (e.g. BMI, a factor found to be important for treatment response prediction in another study) (Iniesta et al., 2016). Our results could also have benefitted from more modalities of data (such as imaging or molecular data), as these pooled models have been shown to sometimes outperform models with fewer data types (Lee et al., 2018). We also used data from strictly controlled clinical trials, combined across many countries that do not mirror a clinical environment, and therefore patient behaviour may differ in the real world (e.g. less strictly monitored medication adherence).

We plan to use this method to predict how MDD symptoms and symptom clusters respond to DVS monotherapy. For example, do patients with more severe somatic symptoms respond more consistently than patients with more severe mood symptoms? Based on patterns of these features, we may investigate whether subtypes of treatment response can be typified, and whether these subtypes of response can be matched to subtypes of depression, and hence to treatment response. It would be interesting to discover how well our algorithms work in other trials and larger datasets, and to discover a classifier with a stable set of features that accurately predicts remission across medications and trials.

Unlike previous work in MDD treatment response prediction, we had the opportunity to train a classifier on data that came from patients in many different health systems, with many different standards of care. Despite these differences, our learned model performed with significantly better-than-chance accuracy on holdout data, which suggests that global, multi-site clinical trial data can be combined for predictive modeling of treatment response.

In addition, we have not yet explored an entire sector of participants in clinical trials: the placebo group. Predicting placebo responders would be a first step toward addressing this confound in clinical trials, and may lead to more effective testing of new medications. This would allow us to remove subjects from treatment response prediction who are likely to have a high placebo response. Machine learners run on a group of patients likely to exhibit low placebo response would be better at predicting medication effects alone, as opposed to our current prediction of medication with placebo effects.

In summary, this machine learning approach is an important step forward for clinical practice, because it demonstrates the feasibility of using easily collected baseline data to improve prediction of antidepressant efficacy. A significant improvement in accuracy of predicting remission over chance confers an advantage to a machine learned strategy over current practice. Applied broadly, machine learned models of treatment prediction may change clinical practice in two ways. First, classification models (such as the one in this study) can identify which patients are likely to remit, for a specified drug. Second, machine learning regression models may allow clinicians to compare remission probabilities of many drugs, towards identifying the best class of drugs (or the best for a given cost, in terms of dollars, or for side effects. These two advantages will help clinicians target both a class of drugs and an individual drug, based on an individual patient's characteristics.

Note regarding clinical dataset selection: one trial (NCT01309542) was open-label with a different dose range than other trials. It is possible that including this trial affected trained classifier accuracy. It will be interesting to explore the relationship of dose, trial type, and classifier performance in future work.

2.5 Funding

2.5.1 Funding source role

Funding sources did not play a role in study design, data analysis, interpretation of results, writing, or submission of this study for publication. All authors had access to all data in the study, and held final responsibility for the decision to submit for publication. Data for this study were provided by Pfizer Canada Inc., as post-capture anonymized data for secondary analysis, use of which was approved by the University of Alberta Health Research Ethics Board - Health Panel, application no. Pro00064974.

Author Contributions

James Benoit, Russ Greiner, Matt Brown, Serdar Dursun, and Andy Greenshaw contributed to the acquisition, analysis, and interpretation of data, and assisted in writing and revising this manuscript.

Bo Cao, Raymond Lam, and Phil Cowen contributed to data interpretation, and in critically revising the manuscript.

Acknowledgments

We would like to thank Pfizer Canada Inc. for their generous provision of the clinical trial data for this study.

Declaration of Interests

SD, AG, RG, PC, and MB have no conflict of interest nor disclosure to make with regards to this paper.

RL has received honoraria for ad hoc speaking or advising/consulting, or received research funds, from: Akili, Allergan, Asia-Pacific Economic Cooperation, BC Leading Edge Foundation, Canadian Institutes of Health Research, Canadian Network for Mood and Anxiety Treatments, Canadian Psychiatric Association, CME Institute, Hansoh, Healthy Minds Canada, Janssen, Lundbeck, Lundbeck Institute, Medscape, Mind.Me, Mitacs, Ontario Brain Institute, Otsuka, Pfizer, St. Jude Medical, University Health Network Foundation, and VGH-UBCH Foundation.

BC is partially supported by the NARSAD Young Investigator award by the Brain & Behavior Research Foundation.

JB received a studentship from Alberta Innovates: Health Solutions to support this work, and has previously received studentship/internship funding from the Natural Sciences and Engineering Research Council, Alberta Innovates: Technology Futures, and Mitacs.

2.6 References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR) [Internet]. 2000. Available from: http://dx.doi.org/10.1176/appi.books.9780890423349
2. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®) [Internet]. American Psychiatric Pub; 2013. 991 p. Available from: https://market.android.com/details?id=book--JivBAAAQBAJ

3. Bagby RM, Ryder AG, Schuller DR, Marshall MB. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? Am J Psychiatry [Internet]. 2004 Dec;161(12):2163–77. Available from: http://dx.doi.org/10.1176/appi.ajp.161.12.2163
4. Brown G, Pocock A, Zhao M-J, Luján M. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. J Mach Learn Res [Internet]. 2012 [cited 2019 Feb 7];13(Jan):27–66. Available from: http://www.jmlr.org/papers/volume13/brown12a/brown12a.pdf
5. Brown MRG, Sidhu GS, Greiner R, Asgarian N, Bastani M, Silverstone PH, et al. ADHD-200 Global Competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. Front Syst Neurosci [Internet]. 2012 Sep 28;6:69. Available from: http://dx.doi.org/10.3389/fnsys.2012.00069
6. Brown TA, Di Nardo PA, Lehman CL, Campbell LA. Reliability of DSM-IV anxiety and mood disorders: implications for the classification of emotional disorders. J Abnorm Psychol. 2001 Feb;110(1):49–58.
7. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry [Internet]. 2016 Mar;3(3):243–50. Available from: http://dx.doi.org/10.1016/S2215-0366(15)00471-X
8. Deecher DC, Beyer CE, Johnston G, Bray J, Shah S, Abou-Gharbia M, et al. Desvenlafaxine succinate: A new serotonin and norepinephrine reuptake inhibitor. J Pharmacol Exp Ther [Internet]. 2006 Aug;318(2):657–65. Available from: http://dx.doi.org/10.1124/jpet.106.103382

9. Ersoy S, Engin VS. Risk factors for polypharmacy in older adults in a primary care setting: a cross-sectional study. Clin Interv Aging [Internet]. 2018 Oct 15;13:2003–11. Available from: http://dx.doi.org/10.2147/CIA.S176329

10. Gartlehner G, Hansen RA, Morgan LC, Thaler K, Lux L, Van Noord M, et al. Comparative benefits and harms of second-generation antidepressants for treating major depressive disorder: an updated meta-analysis. Ann Intern Med [Internet]. 2011 Dec 6;155(11):772–85. Available from: http://dx.doi.org/10.7326/0003-4819-155-11-201112060-00009

11. Gómez D, Rojas A. An Empirical Overview of the No Free Lunch Theorem and Its Effect on Real-World Machine Learning Classification. Neural Comput [Internet]. 2016 Jan;28(1):216–28. Available from: http://dx.doi.org/10.1162/NECO_a_00793

12. Guy W, National Institute of Mental Health (U.S.)., Psychopharmacology Research Branch., Early Clinical Drug Evaluation Program. ECDEU assessment manual for psychopharmacology [Internet]. Rockville, Md.: U.S. Dept. of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, National Institute of Mental Health, Psychopharmacology Research Branch, Division of Extramural Research Programs; 1976. Available from: https://ualberta.worldcat.org/title/ecdeu-assessment-manual-for-psychopharmacology/oclc/2344751

13. Habert J, Katzman MA, Oluboka OJ, McIntyre RS, McIntosh D, MacQueen GM, et al. Functional Recovery in Major Depressive Disorder: Focus on Early Optimized Treatment. Prim Care Companion CNS Disord [Internet]. 2016 Sep 1;18(5). Available from: http://dx.doi.org/10.4088/PCC.15r01926

14. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry [Internet]. 1960 Feb;23:56–62. Available from: https://www.ncbi.nlm.nih.gov/pubmed/14399272

15. Iniesta R, Malki K, Maier W, Rietschel M, Mors O, Hauser J, et al. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. J Psychiatr Res [Internet]. 2016 Jul;78:94–102. Available from: http://dx.doi.org/10.1016/j.jpsychires.2016.03.016

16. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry [Internet]. 2010 Jul;167(7):748–51. Available from: http://dx.doi.org/10.1176/appi.ajp.2010.09091379

17. Jhamb M, Abdel-Kader K, Yabes J, Wang Y, Weisbord SD, Unruh M, et al. Comparison of fatigue, pain and depression in patients with advanced kidney disease and cancer - symptom burden and clusters. J Pain Symptom Manage [Internet]. 2018 Dec 12; Available from: http://dx.doi.org/10.1016/j.jpainsymman.2018.12.006

18. Kennedy SH, Lam RW, McIntyre RS, Tourjman SV, Bhat V, Blier P, et al. Group, CDW, 2016. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 clinical guidelines for the management of adults with major depressive disorder: section 3. Pharmacological treatments. Pharmacological Treatments Can J Psychiatr. 61(9):540–60.

19. Lee Y, Ragguett R-M, Mansur RB, Boutilier JJ, Rosenblat JD, Trevizol A, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. J Affect Disord [Internet]. 2018 Dec 1;241:519–32. Available from: http://dx.doi.org/10.1016/j.jad.2018.08.073

20. McIntyre R, Kennedy S, Bagby RM, Bakish D. Assessing full remission. J Psychiatry Neurosci [Internet]. 2002 Jul;27(4):235–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12174732

21. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry [Internet]. 1979 Apr;134:382–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/444788

22. Oluboka OJ, Katzman MA, Habert J, McIntosh D, MacQueen GM, Milev RV, et al. Functional Recovery in Major Depressive Disorder: Providing Early Optimal Treatment for the Individual Patient. Int J Neuropsychopharmacol [Internet]. 2018 Feb 1;21(2):128–44. Available from: http://dx.doi.org/10.1093/ijnp/pyx081

23. Pearl J. Causality by Judea Pearl [Internet]. Cambridge University Press; 2009 [cited 2019 Feb 10]. Available from: https://www.cambridge.org/core/books/causality/B0046844FAE10CBF274D4ACBDAEB5F5B

24. Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, et al. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. Am J Psychiatry [Internet]. 2013 Jan;170(1):59–70. Available from: http://dx.doi.org/10.1176/appi.ajp.2012.12070999

25. Sheehan DV, Nakagome K, Asami Y, Pappadopulos EA, Boucher M. Restoring function in major depressive disorder: A systematic review. J Affect Disord [Internet]. 2017 Jun;215:299–313. Available from: http://dx.doi.org/10.1016/j.jad.2017.02.029

26. Torres SJ, Nowson CA, Worsley A. Dietary electrolytes are related to mood. Br J Nutr [Internet]. 2008 Nov;100(5):1038–45. Available from: http://dx.doi.org/10.1017/S0007114508959201

27. Trivedi MH, Morris DW, Grannemann BD, Mahadi S. Symptom clusters as predictors of late response to antidepressant treatment. J Clin Psychiatry [Internet]. 2005 Aug;66(8):1064–70. Available from: https://www.ncbi.nlm.nih.gov/pubmed/16086624

28. Trivedi MH, Rush AJ, Wisniewski SR, Nierenberg AA, Warden D, Ritz L, et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. Am J Psychiatry [Internet]. 2006 Jan;163(1):28–40. Available from: http://dx.doi.org/10.1176/appi.ajp.163.1.28

29. U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER). Major Depressive Disorder: Developing Drugs for Treatment, Guidance for Industry [Internet]. 2018. Available from: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM611259.pdf

2.7 Supplementary Materials

Table 2.3. List of 92 features included in the training dataset

Lab Tests	HAM-D	Demographics	
Albumin	Agitation	Age	
Alkaline Phosphatase	Anxiety/Psychic	Sex	
Basophils	Anxiety/Somatic	Ethnicity	
Bilirubin	Depressed mood	American Indian or Alaska Native	
Chloride	Feelings of guilt	Asian	
Cholesterol	Genital symptoms	Black or African American	
Creatinine	Hypochondriasis	Hispanic or Latino	
Eosinophils	Insight	Middle Eastern or North African	
Free T4Z	Insomnia/Early	Native Hawaiian or Other Pacific Islander	
Gamma-glutamyl transferase	Insomnia/Middle	Other	
Glucose	Insomnia/Late	White	
HDL Cholesterol	Loss of weight	Study Location	
Hematocrit	Retardation	Argentina	Poland
Hemoglobin	Somatic symptoms/ Gastrointestinal	Canada	Romania
LDL Cholesterol	Somatic symptoms/ General	Chile	Slovakia
Lymphocytes	Suicide	China	Taiwan
Monocytes	Work and Activities	Colombia	United States
Neutrophils	**MADRS**	Germany	Yugoslavia
Platelet count	Apparent sadness	Estonia	South Africa
Potassium	Concentration difficulties	Finland	
Protein	Inability to feel	France	
Red blood cell count	Inner tension	Croatia	
SGOT (aspartate aminotransferase)	Lassitude	India	
SGPT (alanine aminotransferase)	Pessimistic thoughts	Japan	
Sodium	Reduced appetite	Korea	
Triglycerides	Reduced sleep	Lithuania	
Uric acid	Reported sadness	Latvia	
Urine pH	Suicidal thoughts	Mexico	
Urine specific gravity	**CGI**	**Other**	
White blood cell count	Severity	Polypharmacy count*	

*Count of each reported prescription/non-prescription medication and supplement taken at baseline

2.7.1 Details of the Concure learning algorithm:

As shown in Figure 2.3 (a), Concure takes an input a labeled dataset D (each row describing a patient, and each column a component feature of the patient's description, along with the outcome of Remit or No Remit), and returns a classifier $C_{Concure}$ = Concure(D); that classifier, in turn, takes a description of a patient, and returns the remission label.

First, Concure needs to determine the appropriate subset of features to include. To do this, Concure partitions the labeled training dataset D into 5 disjoint sets of patients $D = D_1 \cup D_2 \cup \cdots \cup D_5$, and sets $D_{-j} = D - D_j$. For each i, Concure fits a Lasso (Least Absolute Shrinkage and Selection Operator) model with Lars (Least Angle Regression) using AIC (Akaike Information Criterion) to D_{-j}. This produces 5 classifiers, each using its own set of features -- here 5 feature subsets $\{ f_1, \ldots, f_5 \}$. Concure computes the intersection of the features of these sets to produce a set of common features, $f^* = f_1 \cap f_2 \ldots \cap f_5$, and then focuses on just these common features within D, which we call D[f^*].

Next, Concure wants to identify a good base learner BL*. It considers the following 11 base learners $\{BL_i\}$ (more information on each learner can be found on the scikit-learn website, scikit-learn.org):

7 stand-alone learners:

- random forest (max tree depth= 5, no bootstrapping, max # features= # input features)
- extra trees (aka **ext**remely **ra**ndomized trees, (max tree depth= 5, no bootstrapping, max # features= # input features)
- k nearest neighbors (neighbors = # input features, uses manhattan distance for the Minkowski metric)
- naive bayes (default parameters)
- decision tree (max tree depth= 5, no bootstrapping, max # features= # input features, uses information gain criteria for measuring split quality)
- (linear) support vector machine (L1 penalty)
- neural net (initial learning rate= 0.0001, 500 iterations max, 3 hidden layers, each of size = ⅔ * # input features)

3 meta-learners that combine a standalone learner in various ways:

- gradient boosting (loss set to 'exponential': recovers AdaBoost algorithm, 2 nodes per tree)
- adaboost (default parameters)
- bagging (random forest base learner)

A voting learner (MM) is also considered that first trains the three individual meta-classifiers shown above, then returns a single trained classifier C_{MM}. When given a novel patient, C_{MM}

returns a label based on the label and confidence in that label given by its 3 trained (meta-) classifiers.

To determine the best base learner, Concure needs to estimate the quality of each base learner $BL_i(.)$. Here, we would like to first run each such learner on the full dataset D, and then evaluate that learned classifier on another test dataset, from the same "target distribution" -- the one that gave rise to D. However, we cannot use D as the test dataset, as the target should be disjoint from the training dataset. So, instead, Concure estimates the quality of applying the base learner $BL_i(.)$ to D, by instead running $BL_i(.)$ on D', where D' is similar to D, then evaluating the resulting base classifier, $BC_i = BL_i(D')$ on a set D" that is similar to D, but (importantly) is disjoint from D'.

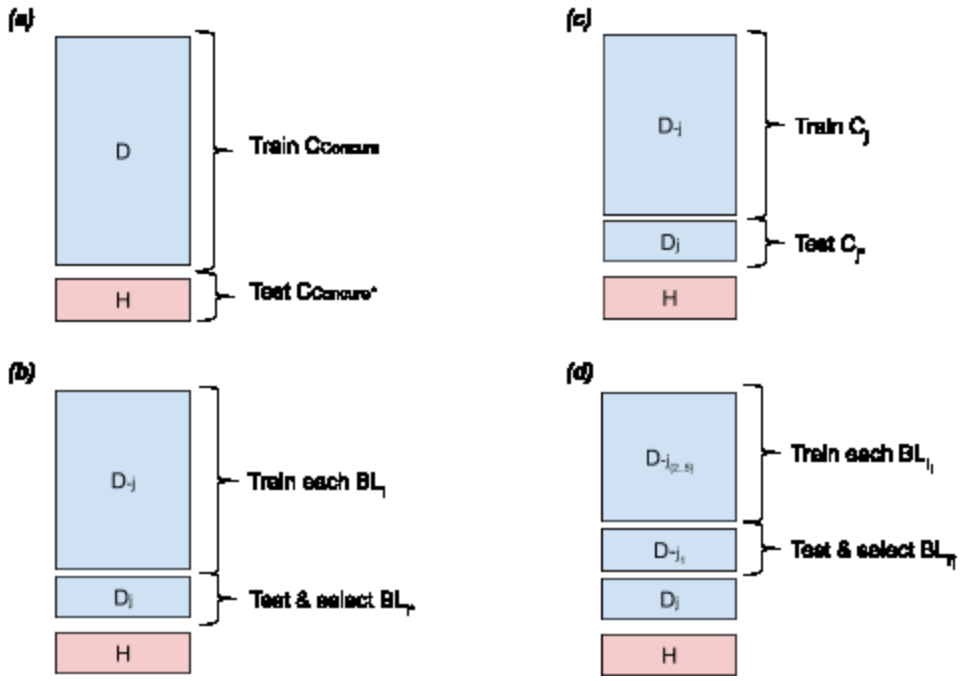

Figure 2.3. Data splitting process for $C_{Concure}$

Figure 2.3 shows the data splitting process for (a) $C_{Concure}$ (training & evaluation of a final trained classifier $C_{Concure^*}$) , (b) BL_{i^*} selection to determine a base learner to be used in $C_{Concure}$, (c) Training of one of five C_{j^*} to provide a meaningful estimate of $C_{Concure}$ quality, and (d) $BL_{i_j{}^*}$ selection to determine which base learner will be used for C_j.

Here, we use 5-fold cross-validation, using the same { D_i } shown above (but here, using only the f^* features for each). For each j = 1..5, Concure runs each of 11 base learners on 4 of the 5 subsets $D_{-j} = D - D_j$; this produces 11 classifiers trained on the f^* projection of D_{-j} -- one for each base learner -- call each $BC_{i,-j}(.)$ for the base learner $BL_i(.)$. It then runs each on the "held-out' fifth subset, D_j -- and uses the results to compute its empirical accuracy:

$$a_{i,j} = acc(\ BC_{i,j}(.), D_j\ ) = \frac{1}{|D_j|}\sum_{(x,y)\in D_j} I[\ y == BC_{i,j}(\ x\ )\ ]$$

where $I[\, y \;==\; BC_{i,j}(\, x\,)\,]$ is 1 if y is equal to $BC_{i,j}(\, x\,)$, and is 0 otherwise.

The cross-validation process actually runs this "train on ⅘, then test on remaining ⅕" process five times, each time holding-out one of the 5 folds. This produces 5 values $\{\, a_{i,1}, a_{i,2}, \ldots, a_{i,5}\,\}$, for each base learner BL_i. Concure then computes the average for each BL_i:

$$s(BL_i) \;=\; \frac{1}{5}\sum_{j=1}^{5} \quad a_{i,j}$$

then claims the best base learner is the one with the highest score $j^* \;=\; argmax_i\{\, s(BL_j\,)\,\}$. Here, Concure(.) found that the SVM base learner had the highest accuracy. (As MM = BL_{11}, j^* = 11.)

Given this, Concure then ran the best base-learner BL_{j^*} = SVM, on $D[f^*]$ (the 26 feature f^* projection of the entire dataset D), to produce our trained classifier $C_{Concure}\;(.)$.

2.7.2 Estimating Predictive Accuracy of Concure Classifier ($C_{Concure}$)

We now want to estimate the predictive accuracy of this resulting $C_{Concure}\;(.)$. We obtain this estimate by running 5-fold cross -validation:

Here, we again divided D into 5 disjoint sets of patients, but now, we ran the entire Concure(.) process on each subset D_{-j}: Concure partitions each D_{-j} into 5 partitions, uses those partitions to find the best feature set ${f^*}_j$, then uses $D_{-j}[\,{f^*}_j]$ to find the best base learner, indexed by ${i_j}^*$,

then runs this base learner BL_{i_j} * on $D_{-j}[f^*_j]$ to produce a classifier, here C_j =Concure(D_{-j}). We then run each of these C_j's on the associated held-out D_j's, to find its accuracy. We then return the average of these accuracies, as our estimate of the quality of Concure(D)'s $C_{Concure}$.

N.b., the 5 feature sets { $f^*_1, \ldots, f^*_5$ }may be different for different j's, and from Concure(D)'s 26-feature f^* , the various base learners selected $\{BL_{1_{j^*}}, \ldots, BL_{5_{j^*}}\}$may be different from each other, and from Concure(D)'s base learner BL_{j^*}= SVM, and the classifiers $\{C_1, \ldots, C_5\}$ may be different from one another, and from Concure(D)'s $C_{Concure}$.

This is irrelevant -- the only reason to generate these 5 $C_j(.)$ classifiers, and then evaluate them on their respective held-out subsets, is just to produce the 5 accuracy values, whose average is used as an estimate of the quality of running Concure on D.

N.b., the Concure(.) learning process is completely automated, and does not involve any human intervention -- in particular, it identified the relevant features, and the best learner, based only on the data it sees.

Table 2.4. Accuracy comparison of classifiers.

		j (fold)					
		1	2	3	4	5	$i_{average}$
i (base learner)	Random Forest	$\mathbf{acc}_{rf,1}$	$\mathbf{acc}_{rf,2}$	$\mathbf{acc}_{rf,3}$	$\mathbf{acc}_{rf,4}$	$\mathbf{acc}_{rf,5}$	$\mathbf{acc}_{rf}$
	Extra Trees	$\mathbf{acc}_{et,1}$	$\mathbf{acc}_{et,2}$	$\mathbf{acc}_{et,3}$	$\mathbf{acc}_{et,4}$	$\mathbf{acc}_{et,5}$	$\mathbf{acc}_{et}$
	Voting (MM)	$\mathbf{acc}_{mm,1}$	$\mathbf{acc}_{mm,2}$	$\mathbf{acc}_{mm,3}$	$\mathbf{acc}_{mm,4}$	$\mathbf{acc}_{mm,5}$	$\mathbf{acc}_{mm}$
	...	...	...	...	...	...	...
	Linear SVM	$\mathbf{acc}_{svm,1}$	$\mathbf{acc}_{svm,2}$	$\mathbf{acc}_{svm,3}$	$\mathbf{acc}_{svm,4}$	$\mathbf{acc}_{svm,5}$	$\mathbf{acc}_{svm}$

One classifier is produced from each combination of base learner and fold. Each classifier produces one accuracy value. The base learner with the highest average accuracy is chosen as i*

Connection Between Chapters 2-3

Remission or non-remission after eight weeks of desvenlafaxine treatment can be predicted with 69.0% accuracy, based on a support vector machine (SVM) classifier trained on data from 3399 patients. The production of this model was accomplished with caveats, both practical and theoretical. Practically speaking, it is difficult for non-domain experts in machine learning to engage with this software in its current form- in order to do so they must first overcome a large knowledge barrier, or collaborate with a domain expert. The broad implications of this barrier are that it leads to a lack of engagement of non-experts in the field. In order to move machine learning into the realm of general clinical use, and drive communication between experts in machine learning and healthcare, the tools being produced (in this case a learned model) must be as interpretable and accessible as possible to all audiences. In addition, experience using machine learning tools will create a greater sense of ownership over these new techniques, decreasing the barriers for adoption and deployment of machine learning tools. As discussed in the next chapter, we attempt to reduce this barrier by introducing automated machine learning software that can be learned and applied to data with minimal effort. It requires an understanding of basic data structure (i.e. each patient case is a row, each clinical feature a column), and the overall objective of using machine learning on clinical data (i.e. predicting something about patients).

In addition, the next chapter considers the role that performance measures play in choosing between algorithms and evaluating classifier performance. While common measures such as accuracy are useful, interpretable performance measures, other measures such as F1 score and ROC-AUC (Receiver Operating Characteristic- Area Under the Curve) may offer information that is not as prone to being affected by, e.g., large differences between patient class sizes.

Chapter 3. Using automated machine learning to predict response in major depressive disorder patients treated with desvenlafaxine.

James RA Benoit, MA[1*], Serdar M Dursun, MD[1], Russell Greiner, PhD[2], Matthew RG Brown, PhD[1], Andrew J Greenshaw, PhD[1]

[1] Department of Psychiatry, University of Alberta, 1E1 Walter Mackenzie Health Sciences Centre, 8440 112 St NW, Edmonton, Alberta, Canada, T6G 2B7

[2] Department of Computing Science, University of Alberta, 2-32 Athabasca Hall, Edmonton, Alberta, Canada, T6G 2E8

[*] Corresponding Author

Abstract

Background

Major depressive disorder (MDD) contributes the most of any disease to the global burden of health, as measured in disability adjusted life years. It has low rates of treatment response, and few tools are available that contribute to treatment planning based on predictive accuracy: selecting an effective antidepressant is primarily determined by trial and error. Patients are often required to switch medications several times before one with an acceptable treatment response profile is found. Algorithms based on machine learning have potential for predicting treatment response; this study uses baseline and two-week data from desvenlafaxine clinical trials for creating machine learning models capable of predicting response or non-response after eight weeks of desvenlafaxine (DVS) treatment.

Methods

We used automated machine learning software, to train machine learning algorithms on data from 2860 MDD patients in 11 phase-III/IV desvenlafaxine clinical trials. Nine classifiers were trained to predict treatment response, defined as an eight-week Hamilton Depression Rating Scale (HAM-D) score of 50% of a patient's baseline score. We trained each model on 60% of the data (n=1716), then evaluated that learned model on a validation set (n=1144), using accuracy as a performance measure. A technique called a cost curve was used to determine whether the same trained classifier should be used for predicting patient response across different patient populations.

Outcomes

The best performing classifier was a trained generalized linear model (GLM), at 70.0% accuracy on a held-out test set, significantly greater than the 54.0% probability of classifying a patient correctly by chance. This model includes six features: CGI-S (Clinical Global Impression-

Severity) score at baseline, and five HAM-D questions taken from two week early response data: anxiety/psychic, feelings of guilt, hypochondriasis, early insomnia, and work/activities.

We demonstrate through cost curves that the GLM is not significantly outperformed by other classifiers tested, suggesting it is sufficient for predicting treatment response across a range of patient populations.

This may allow more accurate use of desvenlafaxine by providing evidence for or against continued treatment response by two weeks after treatment onset. This may contribute to improving psychiatric care through the incorporation of clinical assistive technologies using machine learned models.

Funding

Data for this project were provided by Pfizer Inc. through a data sharing partnership with the University of Alberta.

3.1 Background

Depression has been aptly described as a "blight" on humanity because of the large number of cases (350 million), and years-long duration of disease; half of the world's population lives with negligible access to a psychiatrist for treating this illness (Smith, 2014). Even if every depressed patient in the world were provided with a personal psychiatrist giving them appropriate pharmacological treatment, the number of cases would only drop to 130 million after one treatment attempt: meta-analyses of clinical trials for newer antidepressants found 37% of patients did not achieve a response following treatment for 6-12 weeks (Gartlehner, 2011). While this is a significant drop, albeit using a rather costly solution, it does not address those 130 million non-responders. A path toward improving depression treatment outcomes can be addressed by improving how delivered treatment outcomes can be optimized for efficacy.

Test-retest reliability for many DSM-5 diagnoses is questionable: studies have found less than 25% agreement between interviewers for the diagnosis of MDD in DSM-5 field trials (Regier, 2013). In considering treatment delivery outcomes in depression, there are currently no tests that identify the best treatment for each patient, as we cannot accurately predict a patient's individual response to any antidepressant treatment. As a result, the process of prescribing antidepressants is based on a sequence of trials, until finding one that is effective. While symptom clusters identified in the DSM-5 framework remain a primary information source for diagnosis (APA, 2013), there is no commonly accepted framework or process for utilizing these data in comparing treatment options; e.g., combined low mood and insomnia cannot be used as a biomarker indicating which antidepressant should be prescribed. In addition, the DSM-5 does not incorporate patient genetics, physiology, nor other domains of information that may contribute to

improved diagnostic stratification and treatment recommendation, such as those found in other approaches such as the Research Domain Criteria (RDoC) (Insel et al., 2010).

While we anticipate that treatment outcomes could be improved by considering patients' genetic data, physiology, and other domains of information, doing so requires consideration of three issues related to care delivery: access to these tests is not widespread, the tests require specialized equipment to administer, and the cost of administering these tests is not insignificant; an important consideration in resource-scarce healthcare systems. Treatment delivery could be improved by, e.g., improving access to psychiatric care (e.g. via telehealth), or via the development of protocols that inform nurses and general practitioners which psychiatric treatment is most likely to have an effective, tolerable outcome.

In this study, we develop an algorithm that can predict patient treatment response to desvenlafaxine (DVS) treatment. This algorithm uses data from eleven clinical trials to determine how to use clinical data to distinguish treatment responders from non-responders. Because early response to antidepressants has shown to be a useful predictor of antidepressant efficacy, we included data from HAM-D (Hamilton Depression Rating Scale) (Hamilton, 1960) scores taken at two-weeks after treatment onset (Henkel et al., 2009; Olgiati, 2018). We used these data to train an algorithm called a machine learning classifier on labelled patient data, and test its performance on data where the classifier was blinded to the patient label (see Figure 3.1).

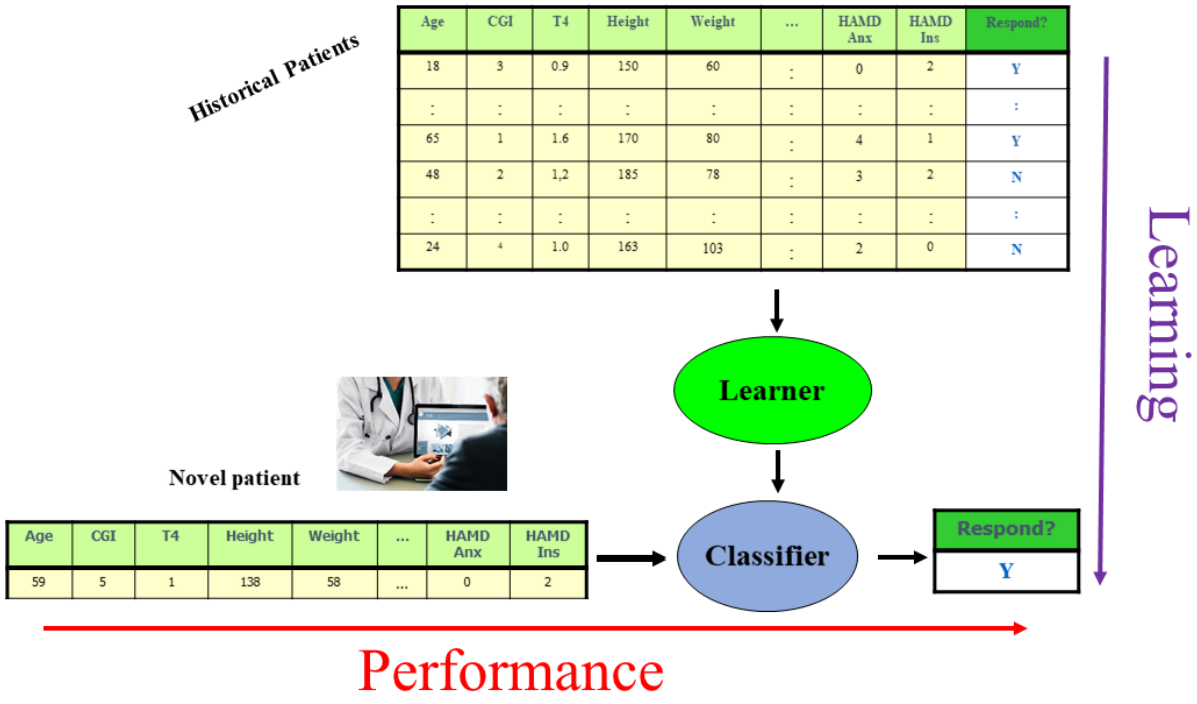

Figure 3.1. Distinguishing between the learning process (top to bottom) and the performance process using the classifier produced by the Learner (left to right).

Machine learning is a subdomain of artificial intelligence focused on creating predictive decision making algorithms. It is becoming a useful technique applied to psychiatry, as it can create accurate models of pharmacotherapy response, using patient information, e.g. easily collected clinical measures such as demographics and Hamilton Depression Rating Scale (HAM-D) items. We diverge from most current machine learning studies by using the RapidMiner machine learning platform to train a machine learning classifier (RapidMiner, 2016). RapidMiner creates an automated pipeline for data preprocessing, producing machine learned classifiers, and generating comparisons between classifiers. Two previous studies have used RapidMiner in the context of psychiatry: an early adopter using it for drug assessments (Kornhuber, 2009), and a recent adopter for assessing MRI images for predicting chronic fatigue syndrome (Sevel, 2018). Similar automated machine learning platforms defined as leaders in the field by Gartner

Analytics' 2019 report: KNIME, TIBCO Software, and SAS (Piatetsky, 2019). These have only been used for non-machine learning studies in the context of psychiatry, focusing on association studies (e.g. Baur et al., 2015). Programs classified as "Challengers" in the field of machine learning, Dataku and Alteryx, remain unused in the literature, based on a PubMed search for each term. Automated machine learning software enables users, who are not experts in machine learning, to use a suite of tools and techniques that would otherwise be inaccessible. The availability of this choice is an important conceptual advance: it shows that integrated machine learning tools are now available to mainstream medical research, and may aid in reducing the knowledge barrier between computing science and clinical practice.

Predicting treatment response can be improved with prior knowledge of a population's response probability and the cost of misclassifying a responder as a non-responder, or vice-versa. Cost is usually framed in terms of the economic cost of misclassification, but can also be viewed in terms of cost to the patient (e.g. in DALY, Disability Adjusted Life Years). However, the specifics of our patient population's cost of misclassification is currently undetermined.
Cost Curves are a technique developed to determine when a classifier's performance is best, and visualize the results better than AUROC (Area Under the Receiver Operating Characteristic) curves. AUROC curves plot classifier performance on a graph with False Positive Rate (FPR) on the X-axis, and True Positive Rate (TPR) on the Y-axis. Each binary classifier tested appears as a point on this graph; the better a classifier is, the lower its FPR and higher its TPR. By contrast, cost curves plot each classifier as a line on a graph (each line is equivalent to an AUROC point; this is called a point-line dual), with Probability Cost on the X-axis, and Normalized Expected Cost on the Y-axis. Probability Cost is given by the formula:

$$PC(a) = \frac{p(a) * C(\overline{a}|a)}{p(+) * C(-|+) + p(-) * C(+|-)}$$

Where a is $-$ or $+$ class, $p(a) * C(\overline{a}|a)$ refers to the misclassification cost of a (composed of the probability that a is the class multiplied by the cost of incorrectly classifying a, and the denominator is the total cost of misclassification (Drummond & Holte, 2006). Normalized Expected Cost is given by the formula:

$$\text{Norm}(\text{E}[Cost]) = \text{FNR} * PC(+) + \text{FPR} * PC(-)$$

where FNR is False Negative Rate, FPR is False Positive Rate, $PC(+)$ is the probability cost of misclassifying a positive case, and $PC(-)$ is the probability cost of misclassifying a negative case (Drummond & Holte, 2006).

In treatment response prediction, cost curves are useful as a means of visualizing trained classifier performance and comparing classifier performance. Here, Cost Curves are used to determine when to use which classifier to make treatment response predictions across the spectrum of probability costs (provided we accept that the definition of "best", here, means the lowest expected cost, a ratio of population composition and misclassification cost) (Drummond & Holte, 2006). It should be noted that different patient populations may incur different costs of misclassification, depending on how misclassification is defined. For example, if we are using DALYs incurred as our definition of cost, a patient responding to depression with predominantly physical symptoms may incur a different misclassification cost than one responding with more mood-based symptoms (e.g. sadness) or motivation-focused symptoms (e.g. anhedonia).

3.2 Methods

3.2.1 Datasets

Clinical data included in this study were drawn from 11 clinical trials of desvenlafaxine. Studies included adult participants ages 18-86, from phase III/IV trials, with HAM-D ≥20 consistent

with criteria used in previous studies (Soares et al., 2014). These data were obtained as part of a data access agreement between the University of Alberta and Pfizer Canada Inc. Approval for this study was given by the University of Alberta Research Ethics Board, study Pro00064974, with all participants providing written consent for their anonymized data to be used prior to analysis being carried out.

3.2.2 Inclusion Criteria

Patient inclusion criteria were a primary MDD diagnosis, completion of treatment with DVS for eight weeks from trial inception, and completion of a 17-item HAM-D assessment at baseline, two weeks, and eight weeks. Treatment response was defined by a participant's eight-week HAM-D score being ≤ 50% of baseline HAM-D score. Patients with comorbid psychiatric diagnoses were excluded. Missing data points (incomplete cases) were filled using mean imputation. In addition, features were excluded when all subjects from one or more studies lacked that feature (e.g. patient weight). After removal of subjects not meeting the inclusion criteria, 2860 subjects and 92 features remained.

Table 3.1. Dataset demographic information and HAM-D17 mean score averages.

n	**2860**
Age (years)	**43.4**
Sex (% Female)	**68.4**
Ethnicity (% White)	**60.3**
HAM-D17 Baseline	**23.5**
HAM-D17 Week 2	**15.8**
HAM-D17 Week 8	**11.5**
Response Rate %	**54.0**

3.2.3 Features considered

We considered scores from baseline lab tests, baseline and two-week HAM-D scores, baseline MADRS (Montgomery-Åsberg Depression Rating Scale) scores (Montgomery & Asberg, 1979), baseline CGI (Clinical Global Impressions scale) scores (Guy, 1976), baseline patient demographics, and baseline medication data in the form of degree of polypharmacy (a sum of the number of reported medications being taken at baseline, including non-prescriptions and supplements). A complete list of these features can be found in Table 3.2.

Table 3.2. List of 109 features included in the training dataset

Lab Tests	HAM-D17 (baseline and week two)	Demographics	
Albumin	Agitation	Age	
Alkaline Phosphatase	Anxiety/Psychic	Sex	
Basophils	Anxiety/Somatic	*Ethnicity*	
Bilirubin	Depressed mood	American Indian or Alaska Native	
Chloride	Feelings of guilt	Asian	
Cholesterol	Genital symptoms	Black or African American	
Creatinine	Hypochondriasis	Hispanic or Latino	
Eosinophils	Insight	Middle Eastern or North African	
Free T4Z	Insomnia/Early	Native Hawaiian or Other Pacific Islander	
Gamma-glutamyl transferase	Insomnia/Middle	Other	
Glucose	Insomnia/Late	White	
HDL Cholesterol	Loss of weight	*Study Location*	
Hematocrit	Retardation	Argentina	Poland
Hemoglobin	Somatic symptoms/ Gastrointestinal	Canada	Romania
LDL Cholesterol	Somatic symptoms/ General	Chile	Slovakia
Lymphocytes	Suicide	China	Taiwan
Monocytes	Work and Activities	Colombia	United States
Neutrophils	**MADRS**	Germany	Yugoslavia
Platelet count	Apparent sadness	Estonia	South Africa
Potassium	Concentration difficulties	Finland	
Protein	Inability to feel	France	
Red blood cell count	Inner tension	Croatia	
SGOT (aspartate aminotransferase)	Lassitude	India	
SGPT (alanine aminotransferase)	Pessimistic thoughts	Japan	
Sodium	Reduced appetite	Korea	
Triglycerides	Reduced sleep	Lithuania	
Uric acid	Reported sadness	Latvia	
Urine pH	Suicidal thoughts	Mexico	
Urine specific gravity	**CGI**	**Other**	
White blood cell count	Severity	Polypharmacy count*	

*Count of each reported prescription/non-prescription medication/supplement being used at baseline

3.2.4 Automated machine learning

3.2.4.1 Data cleaning

After loading the dataset in RapidMiner, we ensured all features were correctly identified by type (e.g. categorical, numerical), used its automated data cleansing option to remove features that were unlikely to contribute to the trained classifiers' accuracy (also called high stability features), and normalized all features to values between 0-1. The cleansing process removed 28 features from the original 109 features, leaving a features space of 81. These 81 features were used to train classifiers to discriminate between two classes: responders and non-responders at eight weeks, as defined previously.

3.2.4.2 Auto Model Process

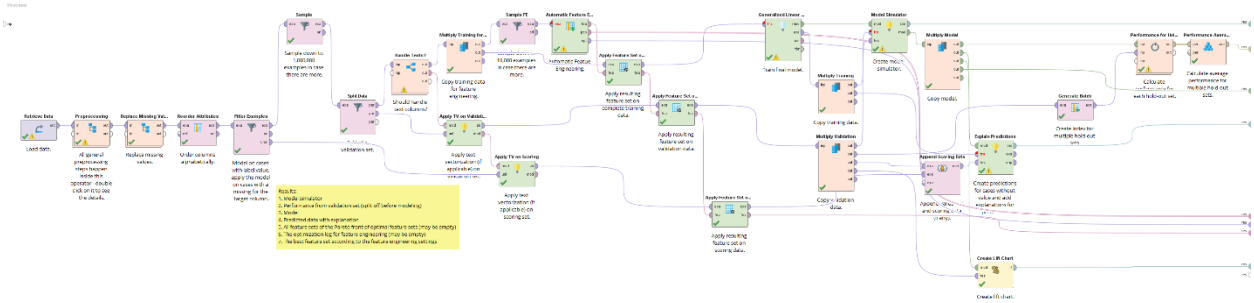

Figure 3.2 RapidMiner's Auto Model process

Figure 3.2 shows RapidMiner's Auto Model process from start to finish. This process works in three basic steps for each classifier produced: data preprocessing, model creation, and performance evaluation. The point of showing this figure is not to read each step; rather, it is to show an overall picture of the machine learning pipeline created by RapidMiner.

Preprocessing steps are carried out on an ad-hoc basis based on the types of data present in the dataset (e.g. numerical, text); this dataset includes missing value mean imputation and alphabetical column reordering.

Once data cleaning and preprocessing are complete, RapidMiner partitions the dataset (n=2860) into primary training (60% of cases, n=1716) and disjoint validation (40% of cases, n=1144) sets. The "feature selection" option was turned on during the classifier training process and set to optimize predictive accuracy over model simplicity. In feature selection, the training set is split again into a secondary training and test set, and RapidMiner repeatedly picks a subset of patient cases and features from the training set, optimizes the weights of those features (i.e. trains a classifier) on that set, and tests the trained classifier's performance on the secondary test set. The classifier with the best performance on the test set is then selected, and the primary training and validation sets are cut to its feature set. Then, the same type of classifier is re-trained on the primary training set, and its performance evaluated on the validation set to produce a final trained classifier.

This process is repeated for each type of classifier produced. RapidMiner's Auto Model process trains nine classifiers, each based on a different type of machine learning classifier. These include:

1. Naive Bayes
2. Generalized Linear Model
3. Logistic Regression
4. Fast Large Margin
5. Deep Learning
6. Decision Tree
7. Random Forest
8. Gradient Boosted Trees

9. Support Vector Machine

Once all classifiers are trained, the test set performance data, confusion matrix, and classification instances are extracted, and performance compared based on one of the available performance measures.

3.2.4.3 Performance measures

The confusion matrix output produced from the test set of each classifier in each experiment was used to calculate accuracy, classification error, F1 score, ROC-AUC, precision, recall, sensitivity, and specificity. Because of the relatively balanced responder/non-responder classes (54.0%/46.0%, respectively), accuracy was used as a primary performance measure. In cases where accuracy was not significantly different between classifiers, model simplicity was compared, and the simplest model chosen.

3.2.5 Cost Curves

A cost curve was constructed to compare the nine trained classifiers produced (See Figure 3.3). Significance testing was carried out on classifiers that were part of the lower envelope (i.e. classifiers with the lowest normalized expected cost at each probability cost) but differed from the most accurate classifier produced by RapidMiner. This was carried out by sampling 1143 cases (with replacement) from the 1143 classification instances of each classifier, creating a new classifier performance line from this sampled data, subtracting one classifier line from the other to produce a line indicating the performance difference between classifiers, and repeating this process 500 times. Then, the highest and lowest 5% of classifier difference lines were discarded, leaving the middle 90% of classifier differences (equivalent to a 90% confidence interval). Ranges on the cost curve where the difference between classifiers was above or below 0 in all instances of the 90% remaining classifier differences are considered to represent significant

differences between the two classifiers (see Figures 3.4 and 3.5). If none of these exist within the cost curve operating range, other classifiers do not show a significant performance difference compared to the most accurate trained classifier.

3.3 Results

3.3.1 Classifier performance comparison

Of the nine trained classifiers produced, a Generalized Linear Model (GLM) classifier was most accurate at 70.05% ± 1.17%. The performance measures for this classifier are summarized in Table 3.3.

Table 3.3. Most accurate classifier performance measures

	Baseline + two-week data
Classifier	**Generalized Linear Model**
Accuracy	**70.05% ± 1.17%**
Classification Error	**29.95% ± 1.17%**
F1 score	**72.17% ± 1.28%**
ROC-AUC	**0.752 ± 0.019**
Precision	**72.18% ± 2.57%**
Recall	**72.25% ± 2.31%**
Sensitivity	**72.25% ± 2.31%**
Specificity	**67.47% ± 4.38%**

3.3.2 Feature Selection

The trained GLM classifier was composed of six features: a baseline CGI score and five HAM-D scores taken at two weeks after treatment began. Table 3.4 summarizes the model produced:

Table 3.4. GLM trained classifier features and coefficients

Feature	Coefficient
Two week Work and Activities	**-0.472554315**
Two week Feelings of guilt	**-0.371111036**
Two week Insomnia/Early	**-0.356185245**
Two week Anxiety/Psychic	**-0.279534481**
Two week Hypochondriasis	**-0.256536738**
Baseline CGI Score	**0.308041628**

A positive feature coefficient indicates that scoring highly on that feature will increase the likelihood that a patient will experience a positive treatment response to desvenlafaxine. The magnitude of each coefficient indicates how sensitive the classifier is to changes in that feature's value.

3.3.3 Cost Curves

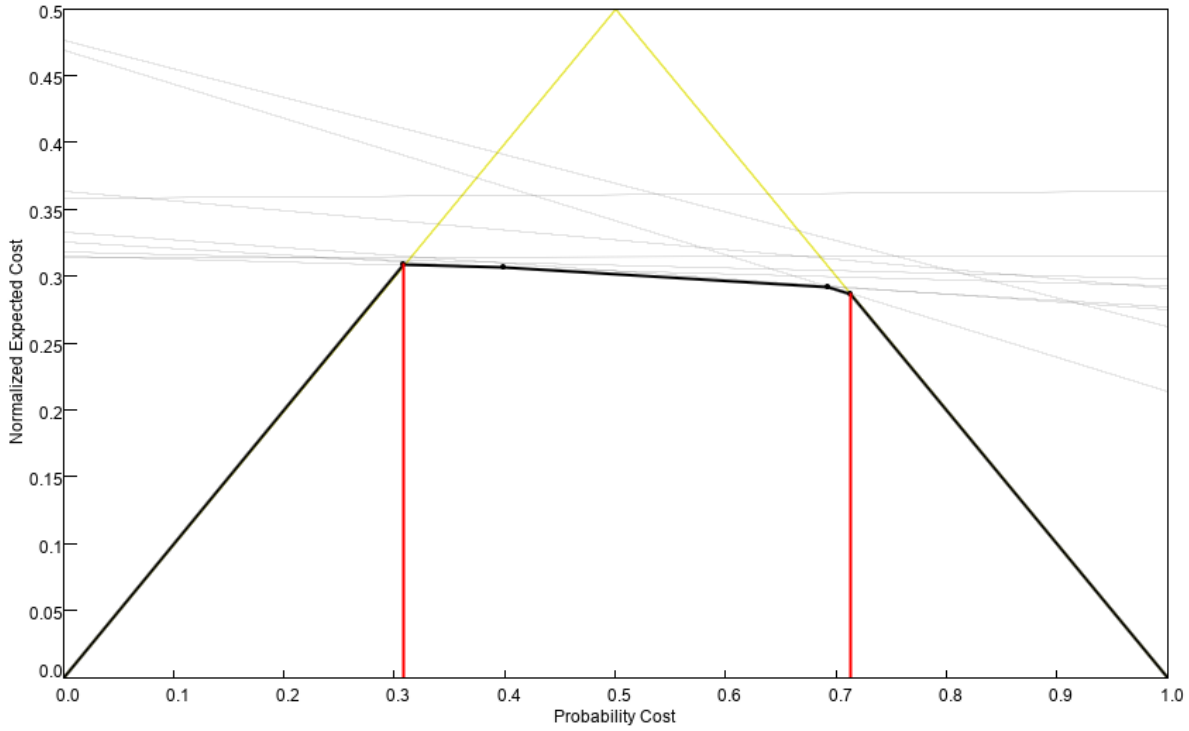

Figure 3.3. Cost Curve comparison of classifier performance

In Figure 3.3, the black line is the lower envelope composed of four classifiers (with each line segment composing it corresponding to one trained classifier), yellow lines are trivial classifiers, red lines indicate non-trivial classifier operating range (where classifier performance was better than trivial classifiers that always respond "yes" or "no" to treatment response), and light grey lines are portions of classifiers not part of the lower envelope. The classifiers composing the lower envelope can be found in Table 3.5. Three classifiers fall within the operating range of the cost curve: GLM, Fast Large Margin (FLM), and Deep Learning. Table 3.5 summarizes the lower envelope:

Table 3.5. Lower envelope composition and boundaries from Figure 3.3

	Lower bound (Probability Cost)	Upper bound (Probability Cost)
FPR 0, TPR 0 (trivial classifier; always guess "No")	**0**	**0.31**
Fast Large Margin	**0.31**	**0.40**
Generalized Linear Model	**0.40**	**0.69**
Deep Learning	**0.69**	**0.71**
FPR 1, TPR 1 (trivial classifier; always guess "Yes")	**0.71**	**1**

Significance testing of the top performing classifier (in this case, GLM) compared to each non-trivial classifier forming the lower envelope, shows that FLM and deep learning do not offer significantly different performance from GLM in the cost curve operating range (Probability Cost 0.31-0.71) (See Figures 3.4 and 3.5, respectively).

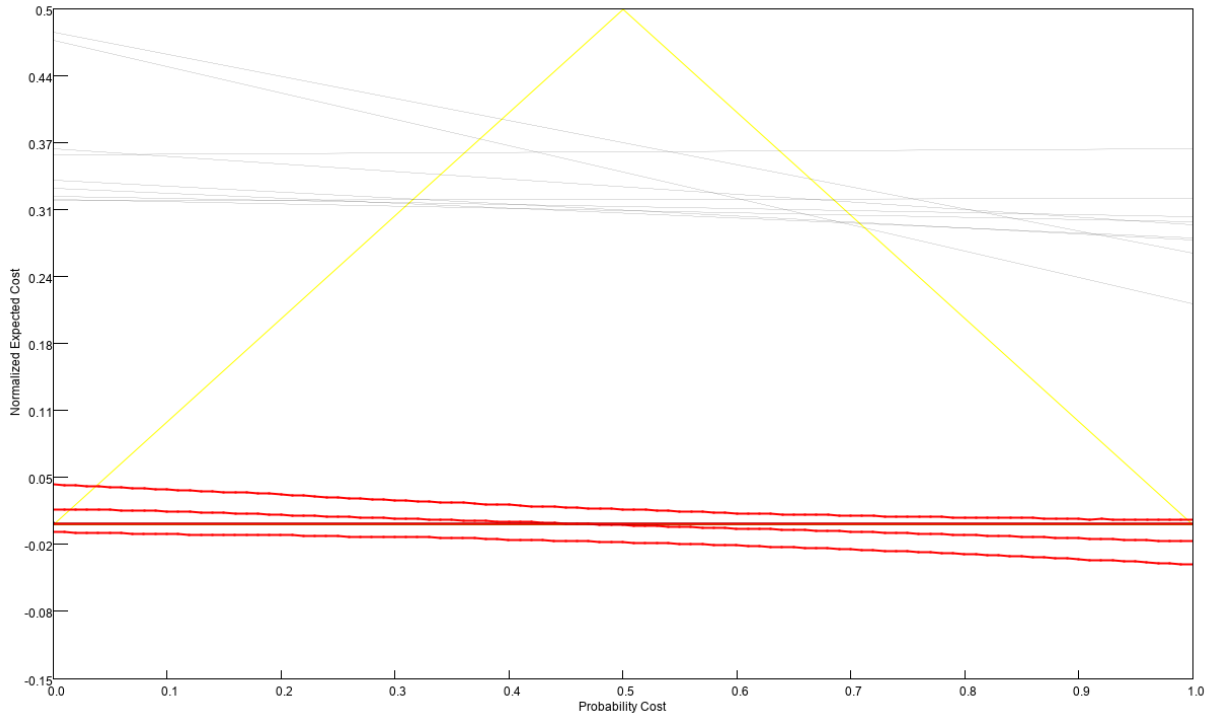

Figure 3.4. Significance testing for trained GLM classifier against trained FLM classifier

In Figure 3.4, The three red diagonal lines indicate the confidence interval for classifier differences; the horizontal red line at y=0 must fall outside this confidence interval to indicate significant differences in classifier performance. The FLM classifier does not outperform the GLM classifier significantly at any point in the curve's 0.31-0.71 Probability Cost operating range.

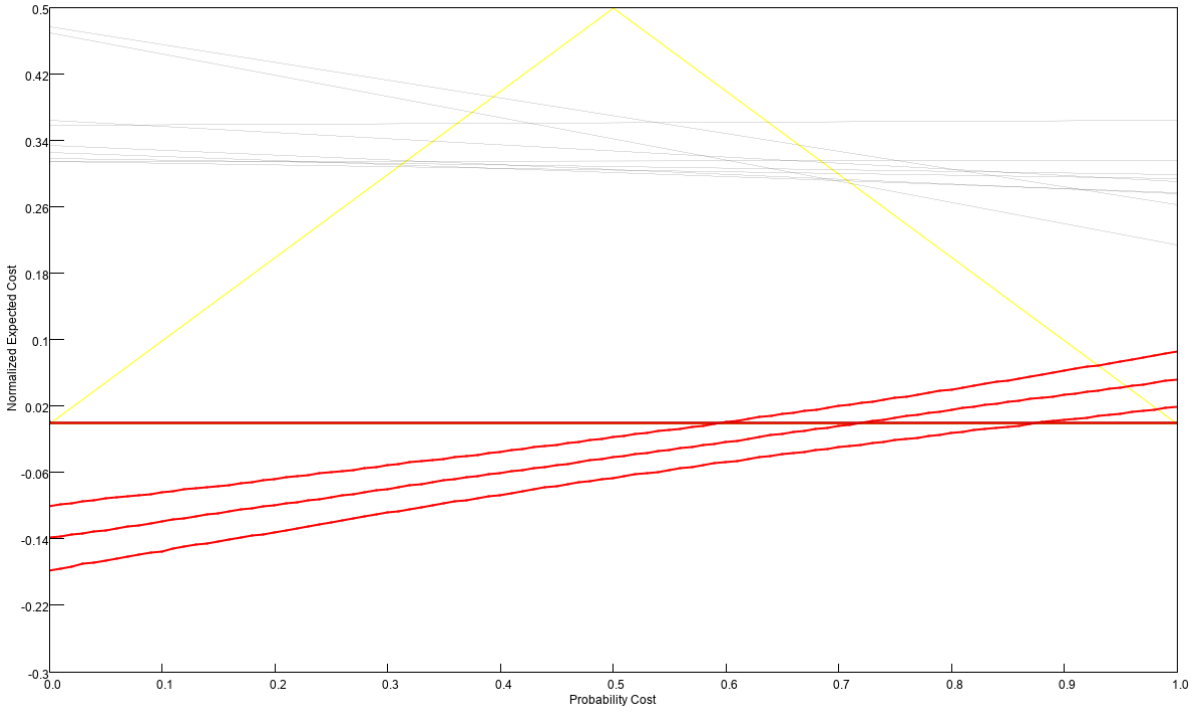

Figure 3.5. Significance testing for trained GLM classifier against trained deep learning classifier

In Figure 3.5, the deep learning classifier does not perform significantly better than the GLM classifier within the cost curve operating range.

3.4 Discussion

Many combinations of confusion matrix outputs are commonly used when considering performance measures with which to evaluate a trained classifier (Wikipedia contributors, 2019), including accuracy, F1 score, or ROC-AUC (Receiver Operating Characteristic- Area Under the Curve). An appropriate performance measure should satisfy three criteria: it should reflect classifier performance over data properties such as severely unbalanced classes, be intuitive to understand, and be the same as similar research in the area, in order to allow for direct comparisons (Straube & Krell, 2014). To this end, accuracy most clearly fills the second and third goals of interpretability and comparability, but its use is dependent on dataset composition. In some unbalanced datasets, accuracy can make classifier performance appear to be high, even if the classifier is simply guessing the majority class; this has been addressed previously by reporting whether classifier accuracy is significantly above the majority class proportion (Chekroud et al., 2016). In the clinical trial data we used, the classes are relatively balanced (54.0% responders, 46.0% non-responders), suggesting accuracy was the preferred performance measure. While the GLM classifier was the most accurate, five of the nine classifiers were not significantly less accurate than it ($p<0.05$). However, the GLM classifier was the simplest, using the fewest features (6 vs 9-15 in the five similar-performing models), and drawing only from questionnaire data (vs at least one of lab test, demographic, or medication data that were included in all other models), further supporting GLM's use.

Only six of the 81 features were included in the GLM classifier: CGI-S (Clinical Global Impression-Severity) score at baseline (Guy, 1976), and five HAM-D features from two-week data collection:

1. Anxiety/psychic

2. Feelings of guilt
3. Hypochondriasis
4. Early insomnia
5. Work and activities

While increased values in CGI are related to an increased likelihood of eight-week treatment response, increased values in any of the five HAM-D features taken at two weeks decrease the likelihood of a treatment response. This suggests that baseline clinical data, alone, is of limited use to predicting eight-week treatment response – it is better to also use the two-week data. We ran a second experiment, excluding two-week HAM-D data, and found the most accurate classifiers were FLM and Logistic Regression, each with 58.90% accuracy. This is an 11.15% drop in accuracy vs the GLM classifier containing two-week HAM-D data. These results support monitoring early response to antidepressants, as well as the creation of clinical tools meant to be deployed at two weeks after treatment onset.

The CGI severity scale is a physician-answered questionnaire consisting of a subjective comparison of the patient's severity of illness to all other patients that clinician has seen. The inclusion of this feature in the GLM classifier supports the importance of the therapeutic alliance, as well as clinician experience in predicting DVS treatment outcomes. The other five HAM-D questions taken at two weeks after treatment onset are drawn from all three dimensions of previous models that attempted to divide HAM-D items into symptom clusters for predicting late response to antidepressants: mood, sleep/psychic anxiety, and somatic anxiety/weight (with hypochondriasis as part of the somatic anxiety/weight cluster) (Trivedi et al., 2005).

Hypochondriasis is not well supported as a feature for treatment outcome prediction: one study suggested that it was one of a constellation of symptoms (including another feature in the GLM classifier, HAM-D Anxiety/Psychic) that predicted depression relapse and recurrence within a two-year period following cognitive therapy (Mallinckrodt et al., 2007). Desvenlafaxine treatment outcomes can be predicted by early response to treatment as measured by HAM-D composite score (Lam et al., 2014), suggesting that lower HAM-D item values at two weeks are associated with better treatment outcomes, consistent with the GLM classifier produced. Anxiety was not predictive of patient response to rTMS (repetitive Transcranial Magnetic Stimulation), although this was based on a 14-item composite score from the Depression, Anxiety, and Stress Scale (DASS), and not the HAM-D, as well as using a different modality of treatment (Lovibond & Lovibond, 1996; Krepel et al., 2019). Patients who responded to treatment with one of four different SSRI's were shown to have lower scores on the Depression and Anxiety Cognition Scale (DACS) than non-responders (Masuda et al., 2017), supporting our trained GLM classifier's results for anxiety. Desvenlafaxine has been shown to have a significant effect on guilt (Kornstein et al., 2009). A composite measure of negative affect consisting of guilt, hostility/irritability, and fear/anxiety items on the HAM-D, has been used to suggest that patients with higher negative affect scores respond better to SSRI (Selective Serotonin Reuptake Inhibitor) treatment (Gerra et al., 2014). However, this measure was not specific to guilt, suggesting it has limited applicability to this study. Insomnia is not supported as a predictor of mono or combination drug therapies in depression, although its presence has been associated with worse treatment outcomes in clinical trials (Sung et al., 2015). This suggests that the feature may have limited applicability in a clinically deployed classifier, and supports the use of clinical data, as opposed to clinical trial data, in future response prediction work aiming to create

deployable classifiers. However, the differences between these data types has not been quantified.

3.4.1 Cost Curve Classifier Evaluation

Cost curves assume that the specific cost of misclassifying a patient as a responder or non-responder, as well as the proportion of responders to non-responders, will be known when the classifier is deployed. Toward this end, they give the useful operating range of each classifier over all possible patient populations, and show the ranges for which each classifier should be used. Here, two trivial classifiers (i.e. always predict "responder" or always predict "non-responder") are present at the extremes of the probability cost spectrum, and three classifiers are present within the produced cost curve operating range. However, the GLM classifier was not significantly outperformed at any point within the cost curve operating range. Taken together with model simplicity, these findings indicate that a trained GLM classifier should be used to predict treatment response unless facing a patient population with an extreme composition of responders or non-responders, or extremely high or low costs of misclassification (or a mixture of both that causes the ratio of costs to fall outside the classifier operating range). However, it should be noted that this finding is representative only of DVS data drawn from clinical trial populations after the application of exclusion criteria that would not be present in general clinical use.

It would be interesting to extend this work to include more types of two-week response data, to determine whether lab tests, demographic data, and other psychiatric scales provide useful features for treatment response prediction. In addition, including four-week response data may provide another increase in treatment outcome prediction accuracy (Olgiati et al., 2018).

3.5 References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Pub; 2013. 991 p.
2. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry. 2016 Mar;3(3):243–50.
3. Drummond C, Holte RC. Cost curves: An improved method for visualizing classifier performance. Mach Learn. 2006 Oct 1;65(1):95–130.
4. Gartlehner G, Hansen RA, Morgan LC, Thaler K, Lux L, Van Noord M, et al. Comparative benefits and harms of second-generation antidepressants for treating major depressive disorder: an updated meta-analysis. Ann Intern Med. 2011 Dec 6;155(11):772–85.
5. Gerra ML, Marchesi C, Amat JA, Blier P, Hellerstein DJ, Stewart JW. Does negative affectivity predict differential response to an SSRI versus a non-SSRI antidepressant? J Clin Psychiatry. 2014 Sep;75(9):e939–44.
6. Guy W, National Institute of Mental Health (U.S.)., Psychopharmacology Research Branch., Early Clinical Drug Evaluation Program. ECDEU assessment manual for psychopharmacology. Rockville, Md.: U.S. Dept. of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, National Institute of Mental Health, Psychopharmacology Research Branch, Division of Extramural Research Programs; 1976.
7. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960 Feb;23:56–62.
8. Henkel V, Seemüller F, Obermeier M, Adli M, Bauer M, Mundt C, et al. Does early improvement triggered by antidepressants predict response/remission? Analysis of data

from a naturalistic study on a large sample of inpatients with major depression. J Affect Disord. 2009 Jun;115(3):439–49.

9. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry. 2010 Jul;167(7):748–51.

10. Kornhuber J, Terfloth L, Bleich S, Wiltfang J, Rupprecht R. Molecular properties of psychopharmacological drugs determining non-competitive inhibition of 5-HT3A receptors. Eur J Med Chem. 2009 Jun;44(6):2667–72.

11. Kornstein SG, Fava M, Jiang Q, Tourian KA. Analysis of depressive symptoms in patients with major depressive disorder treated with desvenlafaxine or placebo. Psychopharmacol Bull. 2009;42(3):21–35.

12. Krepel N, Rush AJ, Iseger TA, Sack AT, Arns M. Can psychological features predict antidepressant response to rTMS? A Discovery-Replication approach. Psychol Med. 2019 Jan 24;1–9.

13. Lam RW, Endicott J, Hsu M-A, Fayyad R, Guico-Pabia C, Boucher M. Predictors of functional improvement in employed adults with major depressive disorder treated with desvenlafaxine. Int Clin Psychopharmacol. 2014 Sep;29(5):239–51.

14. Lovibond SH, Lovibond PF. Manual for the Depression Anxiety Stress Scales. Psychology Foundation of Australia; 1996. 42 p.

15. Mallinckrodt CH, Prakash A, Houston JP, Swindle R, Detke MJ, Fava M. Differential antidepressant symptom efficacy: placebo-controlled comparisons of duloxetine and SSRIs (fluoxetine, paroxetine, escitalopram). Neuropsychobiology. 2007 Nov 23;56(2-3):73–85.

16. Masuda K, Nakanishi M, Okamoto K, Kawashima C, Oshita H, Inoue A, et al. Different functioning of prefrontal cortex predicts treatment response after a selective serotonin reuptake inhibitor treatment in patients with major depression. J Affect Disord. 2017 May;214:44–52.

17. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry. 1979 Apr;134:382–9.

18. Olgiati P, Serretti A, Souery D, Dold M, Kasper S, Montgomery S, et al. Early improvement and response to antidepressant medications in adults with major depressive disorder. Meta-analysis and study of a sample with treatment-resistant depression. J Affect Disord. 2018 Feb;227:777–86.

19. Piatetsky G. Gainers, Losers, and Trends in Gartner 2019 Magic Quadrant for Data Science and Machine Learning Platforms [Internet]. 2019 [cited 2019 Feb 28]. Available from: https://www.kdnuggets.com/2019/02/gartner-2019-mq-data-science-machine-learning-changes.html

20. RapidMiner. Lightning Fast Data Science Platform for Teams | RapidMiner© [Internet]. RapidMiner. RapidMiner; 2016 [cited 2019 Mar 1]. Available from: https://rapidminer.com/

21. Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, et al. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. Am J Psychiatry. 2013 Jan;170(1):59–70.

22. Rumelhart DE, Hinton GE, Williams RJ, et al. Learning representations by back-propagating errors. Cognitive modeling. 1988;5(3):1.

23. Sevel LS, Boissoneault J, Letzen JE, Robinson ME, Staud R. Structural brain changes versus self-report: machine-learning classification of chronic fatigue syndrome patients. Exp Brain Res. 2018 Aug;236(8):2245–53.

24. Smith K. Mental health: a world of depression. Nature. 2014 Nov 13;515(7526):181.

25. Soares CN, Endicott J, Boucher M, Fayyad RS, Guico-Pabia CJ. Predictors of functional response and remission with desvenlafaxine 50 mg/d in patients with major depressive disorder. CNS Spectr. 2014 Dec;19(6):519–27.

26. Straube S, Krell MM. How to evaluate an agent's behavior to infrequent events?-Reliable performance estimation insensitive to class distribution. Front Comput Neurosci. 2014 Apr 10;8:43.

27. Sung SC, Wisniewski SR, Luther JF, Trivedi MH, Rush AJ, COMED Study Team. Pre-treatment insomnia as a predictor of single and combination antidepressant outcomes: a CO-MED report. J Affect Disord. 2015 Mar 15;174:157–64.

28. Trivedi MH, Morris DW, Grannemann BD, Mahadi S. Symptom clusters as predictors of late response to antidepressant treatment. J Clin Psychiatry. 2005 Aug;66(8):1064–70.

29. Wikipedia contributors. Confusion matrix [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Apr 24]. Available from: https://en.wikipedia.org/w/index.php?title=Confusion_matrix&oldid=881721342

3.6 Supplementary Materials

Table 3.6. Desvenlafaxine clinical trial datasets

Dataset	Trial Locations	Year
NCT01309542	Estonia, Finland, Former Serbia and Montenegro, France, Germany, Latvia, Lithuania, Poland, Slovakia, South Africa, United States	2003-2006
NCT00384033	United States	2006-2007
NCT00445679	China, India, Republic of Korea, Taiwan	2007-2009
NCT00406640	Argentina, Chile, Colombia, Mexico, Peru, United States	2006-2008
NCT00369343	United States	2006-2008
NCT00798707	Japan, United States	2008-2010
NCT00863798	United States	2009-2010
NCT01121484	United States	2010-2011
NCT00824291	United States, Canada	2009
NCT00300378	Croatia, Estonia, Finland, France, Latvia, Lithuania, Poland, Romania, Slovakia, South Africa	2006-2007
NCT00277823	United States	2006-2007

Connection Between Chapters 3-4

The viability of machine learning in predicting treatment response was shown by our trained classifier achieving 70.0% accuracy. First, this demonstrated that the RapidMiner automated machine learning software is a viable tool for rapid prototyping of machine learned algorithms, allowing non-domain experts to access ML technology without the steep learning curve that Python-based machine learning entails. Second, it showed the utility of including early response (two-week) data, as it increased predictive accuracy over baseline data alone by 11.15%. However, there is a gap between producing a trained classifier and deploying it in a live healthcare setting. Moving from the production machine learned classifiers, we now turn to the broader field of data-driven personalized medicine and use examples from computational psychiatry to illustrate how machine learning is being applied in the context of the current health environment. This is accomplished through a conceptual review examining four domains of machine learning: access to medical datasets, realigning the expansion of health data with machine learning priorities, machine learning commercialization, and future applications of clinically-oriented machine learning.

Chapter 4. A conceptual review of machine learning in personalized medicine with a focus on psychiatry

James RA Benoit, MA[1*], Serdar M Dursun, MD[1], Matthew RG Brown, PhD[1], Russell Greiner, PhD[2], Andrew J Greenshaw, PhD[1]

[1] Department of Psychiatry, University of Alberta, 1E1 Walter Mackenzie Health Sciences Centre, 8440 112 St NW, Edmonton, Alberta, Canada, T6G 2B7

[2] Department of Computing Science, University of Alberta, 2-32 Athabasca Hall, Edmonton, Alberta, Canada, T6G 2E8

* Corresponding Author

4.1 Introduction

4.1.1 The need for personalized treatment: one size does not fit all

The field of medicine is focusing increasingly on "personalized healthcare". Rather than treat everyone with one disease in the same way, the goal now is to identify (or even generate) the specific treatment that is best for each individual patient, based on the patient's attributes. This has the potential to significantly improve patients' lives and reduce costs by reducing the chance of ineffective treatments. This new direction is being enabled by two recent developments. First is the plethora of extensive databases of electronic medical and electronic health records, describing many aspects of large sets of previous individual patients: symptoms, histological reports, images, and now, the wave of "omics" data (Heart, Ben-Assuli & Shabtai, 2017). Importantly, many of these datasets also identify the "outcome" of whether the patient had a specific disease, or how well said patient responded to a specific treatment. The second enabler is the development of powerful tools (many from machine learning) that can use this information about earlier people to produce systems capable of making accurate predictions about novel patients — for example, diagnosis of a certain disease, or whether a specific patient will respond well to a certain treatment (Shickel et al., 2018).

4.1.2 Differentiating standard (bio)statistics vs supervised ML

There are many ways to analyze a dataset. The prevailing approach is to seek biomarkers: features that, individually, are correlated with the outcome. This approach has proven effective at helping to understand the underlying etiology of the disease, and identifying which test to run next, to better understand the disease (Kalia & Silva, 2015). However, this approach was not designed to diagnose a new patient, nor is it capable of adapting to patients expressing the same symptoms with different underlying causes. In psychiatry, there are no accepted biomarkers for schizophrenia, major depressive disorders, and bipolar disorders, or any other category of

psychiatric disorder (Lozupone et al., 2019). In order to understand the types of questions in personalized medicine that are well addressed with a machine learning approach, as well as its caveats, it will be useful to briefly consider the history of the field.

4.1.3 History

Machine learning and the development of modern computing have developed together since Turing's test in 1950 (i.e. whether a computer could appear indistinguishable from a human) (Copeland, 2004) and Grace Hopper's contributions to English-language programming in 1952 (Wikipedia Contributors, 2019). Less than ten years later, Arthur Samuel, who coined the term "machine learning," was able to demonstrate that an organized collection of wire and vacuum tubes (aka the IBM model 701) could beat a human at checkers (Wikipedia Contributors, 2019). Further developments saw computers increasingly capable: identifying objects with an artificial eye (Rosenblatt, 1958), navigating a cart through an obstacle course (Moravec, 1983), and learning to speak (Sejnowski & Rosenberg, 1987). However, the unrivaled promise of AI (Artificial Intelligence) in 1967 went undelivered: that,

"Within a generation ... the problem of creating 'artificial intelligence' will substantially be solved" -Marvin Minsky (Minsky, 1967)

The first AI Winter occurred when cutbacks to DARPA funding in the US and a damning expert report in the UK resulted in a lack of funding support for AI research in the 1970's. The ironic failure of commercial AI technology to adapt to the market, and repeated failures to develop a general purpose robot, resulted in a second AI Winter, beginning in the mid 1980's (Wikipedia Contributors, 2019). IBM's 1997 resurrection of machine learning as a chess program capable of beating the best human players (Kasparov vs Deep Blue), heralded a new era for artificial intelligence: solving specific problems as part of a larger system, rather than attempting general

purpose solutions. Medical applications of expert systems that lay the groundwork for AI began in the 1970's with Mycin (a system for identifying infection etiology and recommending a dose-adjusted antibiotic) (Shortliffe & Buchanan, 1975); these systems are now common in cardiology for EKG automated diagnosis and CHD risk assessment, and in radiology for X-ray interpretation (Deo, 2015). In all of these examples, an artificial system is learning how to complete a task using information. As a result of this learning process, systems that learn a pattern in the information successfully are able to complete that task.

4.1.4 Current use

One branch of machine learning is called supervised machine learning, where a learning algorithm is given a labeled dataset. This type of dataset may describe a set of people, along with a "label" (typically case vs control), where each person is described using various features: clinical, or genomics data or MRI scans (or combinations thereof). In contrast to machine learning, the traditional bio-statistics approach often involves finding "biomarkers"— identifying which single gene (or individual metabolite or brain region) is most associated with a disease phenotype. While these univariate associations can be extremely useful for understanding the disease itself, and perhaps identifying which further experiment to perform (e.g., which gene to knock out), they do not, by themselves, necessarily determine the diagnosis nor the best treatment for a specific patient, as that typically requires determining the set of features that collectively predict the phenotype. ML provides technologies that can find such combinations of features from earlier data. In contrast, AI is a more general term that covers the development of human-like capabilities in machine form (e.g. creating vision processing systems).

4.1.5 Domains covered by this review

Data access: Medical datasets are growing exponentially in their size, complexity, and availability. Among these will be the eventual application of Electronic Medical Records (EMRs) into current practice, towards an improved standard of care. Awareness of the need for health data standardization for machine learning tasks, as well as consideration of the ethical implications status of health data interacting with newer technologies (e.g. streaming data, data ownership), is important as healthcare becomes a decentralized service.

Movement away from traditional statistical models: There is a clear need to realign the expansion of health data with machine learning as we prepare to move away from smaller samples of the population and toward true population-sized datasets. It is becoming important to review and develop objective measures of data (measures created based on objects, e.g. using neuroimaging features and/or human-administered psychiatric scales for diagnosing depression) that are also useful to machine learning prior to one measure becoming the standard for use. This is especially important for areas such as biomarker development, which has faced issues of reproducibility.

KT/commercialization: Moving into a population-level era of streaming data while maintaining rapid progress on-par with new technology will necessitate a move towards industry partnerships. However, there are clear concerns surrounding trust and certification of commercial applications, especially with sensitive data. Here, a key question is how an environment of trust can be created and maintained between patients and commercial entities, while developing a health implementation environment that will not stifle innovation.

Futurism: Finally, we examine how machine learning can be used creatively in the clinic. We present a number of present-day examples, and consider possible tools such as an AI-based consultant capable of understanding and contributing to health care team meetings.

4.2 Data Access

4.2.1 Practicality of data collection & data management

4.2.1.1 Dataset Accessibility

Access to datasets is becoming a mainstream priority in technology development, as evidenced by the deployment of tools that provide researchers with a unified search platform such as Google's Dataset Search (Dataset Search, 2019). The availability of resources to store and work with extremely large datasets is also rapidly becoming a reality due to the expansion of Cloud computing, the use of storage and computational resources on the internet (Iniesta et al., 2016). With these advances in access, datasets are being used to generate new knowledge, disseminate knowledge faster, translate personalized medicine into practice, and empower patients by enabling easier access to their medical information (Murdoch & Detsky, 2013). For psychiatrists, machine learning programs will be able to search through long natural language records of patient visits from, e.g. clinician notes, and finding patterns of language. Accessibility of the free-text medical record will be crucial for the success of automated clinical decision support tools (Sittig et al., 2008). Current work in this area includes creating systems for accessing patients' personal health records (PHRs) that integrate blockchain technology for record security and the capability to pull data from different health providers in order to create a single unified PHR (Roehrs, da Costa, & da Rosa Righi, 2017).

4.2.1.2 Machine Learning Tool Accessibility

Domain expertise in machine learning is slowly being phased out as a requirement of applying machine learning tools in research. Improvements to usability and automated machine learning pipelines in commercially available machine learning packages such as Rapidminer, Alteryx, and KNIME, are creating environments nearing readiness for use by researchers. The Gartner Analytics Magic Quadrant report is one metric that rates these tools by their completeness of vision (composed of, e.g., product strategy and innovativeness), and ability to execute (composed of, e.g., user experience and product functionality) (Gartner Reprint, 2018; Piatetsky 2019).

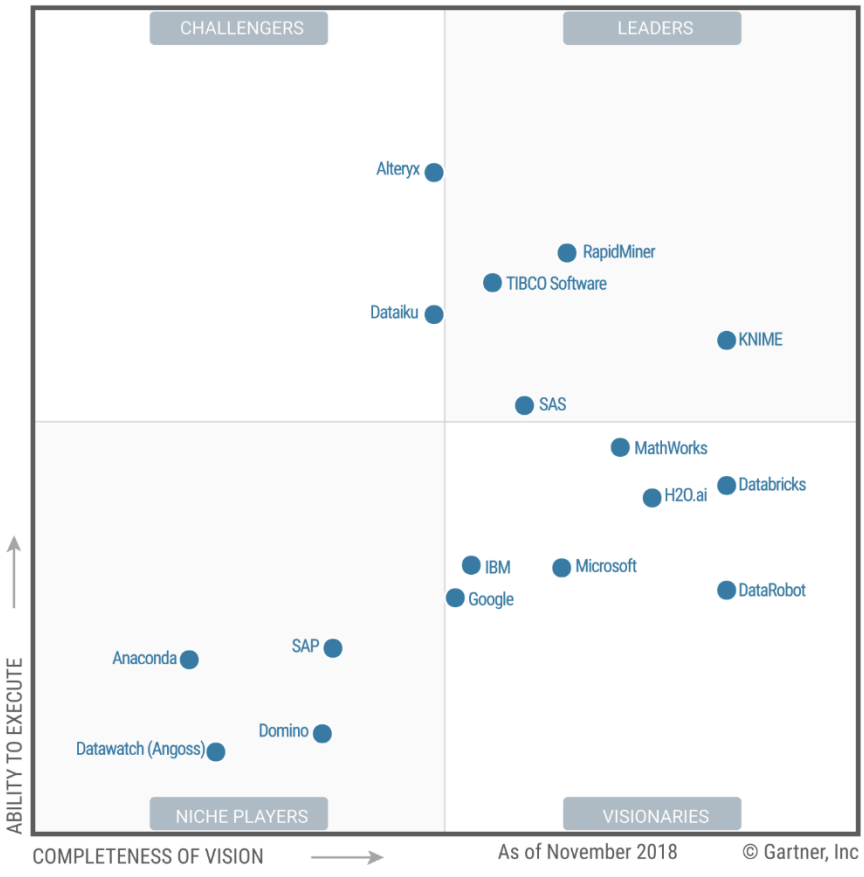

Figure 4.1. Reproduction of Magic Quadrant for Data Science and Machine Learning Platforms

Figure 4.1 is reproduced from from https://www.gartner.com/doc/reprints?id=1-65WC0O1&ct=190128&st=sb.

However, these tools still lack the level of rigour required by medical machine learning. Exporting tool performance measures (e.g. Receiver Operating Characteristic (ROC) curves), is not possible with some tools (e.g. RapidMiner). End-users must therefore rely on domain experts to create secondary programs enabling capture or production of the fine-grained data required for study reproducibility.

4.2.2 Problems of data standardization

4.2.2.1 What is being collected/stored/shared (and is available for analysis)

Healthcare data being used for machine learning appears to have a limited lifetime of utility for some predictive tasks, decaying in usefulness by half for every 4 months of age due to changing clinical practice patterns (Chen et al., 2017). One of the primary heuristics for deciding on whether data is useful to collect for machine learning in medical applications is whether the prediction to be made with that data would already be obvious to a clinician (Chen & Asch, 2017). While many disciplines of medicine are fielding machine learning results with high reported accuracies, these results detract from more difficult-to-approach problems by providing a basis for too-hasty generalizations. For example, if two research questions are asked in the same domain (e.g. health research), with similar data inputs (e.g. a hip MRI vs. a brain MRI), a common assumption is that machine learning performance will be similar between the two. This is an important assumption to address, toward preventing what Chen & Asch (2017) refer to as a "trough of disillusionment," one of the factors that precipitated falling into an AI winter, as previously discussed.

One issue for data collection is multi-source data integration. Medical data exists in at least 10 different modalities (e.g. clinical trials, clinical registries, biometric sensor data), but is often

collected using different scales (e.g. lab tests in ng/mL vs g/L), is missing information, or is incongruent between modalities (Lee & Yoon, 2017).

4.2.2.2 Are we collecting the right data points (how is data being collected)

In EMR development, physicians feel there is a tension between standardized form entry and entering text freely: while entering data into a common set of fields contributes to EMR utility, free text enables nuanced entries that do not lose situational context (Terry et al., 2014). This is an important distinction for machine learning, related to Natural Language Processing (NLP) – a field that develops computers' abilities to understand human language without specific input structures (e.g. single-word entry forms). Computers analyzing patient records need to be able to adapt to an uncommon medical lexicon not used in standard speech or writing.

In drug development, machine learning can be applied to detect relative sensitivities of different modalities of data to the effects of a new treatment, enabling removal of less sensitive data modalities and informed decisions, maximizing data contribution to data cost ratios and ensuring the most salient data is being collected (Doyle, Mehta & Brammer, 2015).

Missing data is also a concern for large datasets, and a three-case taxonomy for missing data has been suggested: data can be missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) (Lee & Yoon, 2017). MCAR data has the same probability of being missing across all subject cases and variables, while MAR data can be traced to observed data points (e.g. certain patients dropping out of a study), and NMAR data is dependent on unobserved data points (e.g. 50% of the values for a given feature are not recorded).

4.2.2.3 Issues collecting from multiple sites: batch effects, covariate shift, and domain shift

The benefits of including more subjects in a learning model are diminished by batch effects: error produced by biases in data collection between groups of subjects resulting from, e.g., different locations, equipment, or group demographics. Correcting for batch effects is difficult, but possible: an empirical Bayes method called ComBat has been used successfully for correcting epigenetics data in a machine learning study of MDD (Malki et al., 2016). MRI scanners are especially susceptible: two identical scanners at the same location, scanning the same patient, will produce different images (Kostro et al., 2014). These effects can be reduced through the application of cross-modality software tools. Previous work has shown the efficacy of using genomics batch correction software using ComBat, for the correction of DTI images in studies looking at psychiatric illness vulnerability, and autism (Fortin et al., 2017).

Similar to sample selection bias (in which data is selected non-randomly for inclusion in an analysis), covariate shift is a case in learning problems where the training set distribution is different than the testing set distribution (Bickel, Brückner & Scheffer, 2009). This difference has been shown to occur in observational studies of treatment effects, due to physicians assigning treatment protocols based on disease severity. Using these biased observations of treatment effects is problematic when, e.g. estimating antidepressant treatment effects (Wen, Hassanpour & Greiner, 2016).

As evidenced by changes in the DSM between versions, the nature of how mental health disorders are measured changes over time. While it is unlikely that humans have evolved new neurochemical processes and accompanying deficits since 1952, the way we measure the same underlying disorders has changed. Domain shift in psychiatry refers to changing metrics for measuring disease processes (Quiñonero-Candela, 2009), and can occur both longitudinally and cross-sectionally. Heterogeneity of measurement methods across study sites produces a dataset

that requires one of two equally poor solutions: change part of the data to mimic the feature or label of the other site, or change both sites' data to fit a hybrid model of measurement between metrics. Making each metric its own variable in this case is a third possible solution, provided the learning algorithm can successfully cope with missing data.

4.2.3 Ethics of Data

4.2.3.1 Impact of using personal healthcare data

Kaelber et al (2008) bring up an interesting topic related to the impact of personal healthcare data: who controls access to a child's PHR (personal health record), and should children be allowed to access their own data (Kaelber et al., 2008)? Considering this case through the lens of ethical jurisprudence reveals a number of issues. For example, when children reach their age of majority (ignoring issues such as whether that age is tied to the individual or can change with location, and whether it applies to decentralized data): who owns their PHR entries? If the parent or guardian has been keeping records on behalf of their child, is there an obligation from the incipient adult to curate their own records? And what is the appropriate balance of beneficence if the impact of entering data will eventually cause negative consequences (e.g. denial of insurance) to the child, but is immediately useful to their wellbeing? One solution for adolescents is to be given graduated access to their records, such that sensitive conditions are kept private while conditions requiring parental involvement are shared (Sittig & Singh, 2011).

This becomes an important issue as we move towards the use of healthcare apps for paediatrics such as AddressHealth (http://addresshealth.in/web/). There is already significant controversy surrounding newborn screening (NBS) programs, which can lack consent and continued-use regulation (O'Doherty et al., 2016).

4.2.3.2 Data storage and accessibility by PHR

If data is irrelevant to treatment and wellbeing of the patient, should it be kept for posterity, and at what point do we have enough evidence that stockpiling data will result in diminishing returns to the standard of care? There is certainly interest in provisioning PHR access tools: 94% of postpartum women expressed interest in PHR access (Fernandez et al., 2017), and at the same time key early EHR adopters such as Kaiser Permanente Northwest are driving this process forward. Another important issue for the volume of data collected is equity between urban and rural patients: internet access, PHR management, and contact with healthcare providers is all decreased in rural areas (Greenberg et al., 2018). In psychiatry, this disparity may be somewhat offset by differences in need: urban dwellers are more likely to develop mental illness (Vassos et al., 2016).

4.2.3.3 Data privacy

Data privacy should be discussed in a sliding context when considering how to prevent data being attributed to individuals when moving from an EMR to EHR environment. The identifiability of subjects using information derived from, e.g., DNA markers or microbiome profiles (Jensen, Jensen & Brunak, 2012), will become a greater concern as tools for reverse-engineering these data to identify individuals in the study proliferate and improve. However, these concerns will not necessarily impact patient choices of which health data they choose to share. A qualitative study examining patient perspectives on sharing anonymized data suggested that although there were concerns about data being shared inappropriately, the vast majority (98%) of patients accepted these risks after considering the perceived altruistic benefit of sharing their data (Spencer et al., 2016).

Since historical data is stable over time, the cumulative probability of it being accessed inappropriately will increase unless a strategy for long-term management is implemented. These strategies potentially include two types of approaches: bottom-up and top-down.

In addition to EMRs, there are now thousands of mobile health (mHealth) apps, of which at least 24 are used for mobile Personal Health Record (mPHR) access (Zapata et al., 2014). The primary concerns with these apps center around practical issues: privacy policy accessibility, ensuring only the intended user is accessing the data, and whether the app follows security standards (e.g. conformity to HIPAA (the U.S. Health Insurance Portability and Accountability Act of 1996). Similarly, within the mental health app ecosystem, privacy is the first and most pressing concern. This is followed by a hierarchy of app efficacy, user engagement, and data sharing, respectively (APA, 2019). This hierarchical framework suggests that apps intended for clinical use should not continue to be considered for use if they fail to meet the standards of use at each stage of assessment (Torous et al., 2018).

4.2.3.4 Bottom-up strategy for data privacy

Blockchains (Fortney, 2019) are digital ledgers, with a few special properties that make them extremely difficult to corrupt (Wikipedia contributors, 2019). They enable new options to keep health data secure and uncorrupted without requiring the time to encrypt and transfer an extremely large dataset. For example, storing EMR data in a hospital, but keeping hashes (dictionaries for the data) in a blockchain, would enable efficient retrieval of records (Esposito et al., 2018). For machine learning purposes, this necessitates moving the algorithms to the stored data, which is a much more computationally palatable option as the algorithms are generally small and do not require encryption to move from place to place.

An alternative approach to blockchain has been suggested, in order to emphasize patient control over data: adding DRM (Digital Rights Management)-like features to EHR data to prevent negligent storage and inappropriate movement (Jafari, Safavi-Naini & Sheppard, 2011). This approach suggests that patients own their own data, but can issue licenses for access and use. This presents potential problems for the machine learning community, since access and use are required for development of new tools. In addition, it risks skewing all datasets by only including subjects who are willing to share their data. This has implications for tool generalizability, and may limit the development of optimal, population-level machine learning solutions.

Maintaining transparency of data is key to gaining public acceptance to data sharing- the failed care.data initiative in the UK is a good example of how a failure in public trust over sharing data with researchers and businesses has the potential to derail healthcare data sharing initiatives (Kostkova et al., 2016). One solution proposed by Kostkova et al. (2016) is to involve the public in shared goal setting to determine data sharing core principles, and as part of this process include mechanisms for strong disclosure and notification systems for data sharing violations (Kostkova et al., 2016).

4.2.3.5 Top-down strategy for data privacy

This strategy mandates that health data not be used in areas where health-based discrimination from algorithmic bias could occur (Hajian, Bonchi & Castillo, 2016), such as insurance providers using it as part of a client risk assessment. One solution may be to enact policies that allow individuals with objections to sharing data to be able to opt out of data sharing. However, this point merits consideration of the detriments to society in allowing data control: how do we decide the consequences of and boundaries to limiting access to data (Kaplan, 2016)? Clinicians have a common concern about how data they enter into an EMR will be shared, centered around

unknowns of who will be able to access data, especially relating to secondary access: for example, should researchers doing secondary analysis get the same level of access as primary care providers (Terry et al., 2014)? In a similar thread, having data available on any centralized online system increases the damage potential from a breach. This is an important consideration in light of the 2017 Equifax data breach that compromised the name, birthday, and social security number of almost half of U.S. citizens (Gressin, 2017).

4.3 Movement away from traditional statistical models

4.3.1 Model-driven vs data-driven approach

4.3.1.1 Population vs sample data

As more health authorities switch to an EHR-based system, data will be curated and available for machine learning at a population level. When data becomes available on this scale, the predictive models used can be simplified, data no longer requires expert annotation, and missing data points can potentially (depending on why the values are missing) be effectively filled in from a corpus of data with millions of examples (Halevy, Norvig & Pereira, 2009). As a result, these data can allow for a data-driven approach.

4.3.2 Predictive vs associative modeling

4.3.2.1 Biomarker discovery (long term) vs. tools to use with patients (short term)

An important distinction between prediction and association studies is the type of finding: biomarker discovery (long term) vs. tools to use with patients (short term). Association studies attempt to explain a biological process by seeking the best (most highly correlated) features in a dataset, while prediction studies seek a model, based on a set of features capable of predicting the class label of new patients. This set of features is sufficient for prediction tasks, but is not necessary or causal- it does not explain the process at hand. Therefore, if finding biomarkers to explain a process is the goal of a study, an association study is warranted, while a study focusing

on making the best medical decisions will be better served by a prediction (machine learning) study.

4.3.2.2 New drug development is time- and cost-prohibitive

The out-of-pocket cost for a new drug is over $1.3 billion as of 2013 (DiMasi, Grabowski & Hansen 2016), and for some fields, have less than 1% chance of success (e.g. Alzheimer's drugs have a 99.6% failure rate since 2002) (Cummings, Morstorf & Zhong, 2014). An associative modeling approach to this problem relies on basic science to develop biomarkers and disease models in order to create principled approaches to new therapy development. Importantly, as these approaches become more complete and nuanced, we could develop models that (also) predict adverse effects, facilitating the development of drugs that would have decreased treatment non-adherence and discontinuation (Bull et al., 2002).

4.3.3 Evaluating Outcomes

4.3.3.1 Biases in data

Machine learning tasks begin by defining a performance task. This identifies the population of interest, range of desired outcomes, and evaluation criteria. Data can then be collected relevant to the performance task. However, some elements of performance tasks can be unknown (e.g. we are predicting treatment response in a new or poorly defined disease).

Datasets from clinical trials and research studies are rarely equivalent to the real-world setting containing the disease/process they attempt to model. While limiting the number of variables creates a good testing environment, machine learning performed on this data will not account for variance removed for the purpose of creating a better disease model. To attempt to account for this discrepancy during the learning process, a number of techniques can be applied. First, the costs can be balanced between groups by applying a cost-sensitive learning function. This ensures a

classifier does not simply learn to guess the majority class. Second, we can reframe a classification problem using severely unbalanced groups in terms of anomaly detection rather than classification, allowing different techniques to be applied to the data (Soni, 2018).

4.3.3.2 Progress in outcome evaluation

There are many assessments for evaluating classifier performance: a few commonly used in evaluation of medical machine learning tools are accuracy, AUROC score, and sensitivity/specificity. However, these scores and significance tests fail to incorporate the varying probabilities of seeing a positive case, against the cost of misidentifying that case. This is an important distinction for personalized medicine, because it would allow us to break down the operational range of tools based on an assessment of a condition's rarity against the cost of getting a test wrong. In psychiatry, for example, there are distributions of how well patients respond to an antidepressant, and the cost of misidentifying the treatment as effective will vary with their degree of depression. In this case, two patients with the same condition may need to be assessed differently: not just in terms of whether they should take a drug or not, but in terms of which scale (or machine learned tool) should be used for assessment. The question of validity is important here: current "gold standard" scales such as the HAM-D (Hamilton Depression Rating Scale), may only be the best test within a certain range of symptom severities. A related problem is dose management, especially salient in heterogeneous areas (e.g. urban zones) where patient populations with diverse ethnic backgrounds show significant variance in CYP2D6 expression (McLellan et al., 1997), which will have different costs associated with misclassification based on the drug given: some drugs have more serious misdosing effects than others. Solutions to this ambiguity involve using an assessment tool called a Cost Curve, which assesses classifiers across

a broad range of population compositions and misclassification costs (Drummond & Holte, 2000; Drummond & Holte, 2004; Drummond & Holte, 2006).

4.3.4 Which subset of clinical questions benefit from a ML approach?

4.3.4.1 Use in Screening, Diagnosis, Prognosis

Machine learning is set to disrupt current medical practice in three areas: improving prognostic models, outdoing human medical image interpretation, and reducing diagnostic errors (Obermeyer & Emanuel, 2016). With recent advances in deep learning, health event prediction across multiple centers using EHRs is now possible, even without harmonization of data between centers (Rajkomar et al., 2018). This technique is especially accurate at predicting within-hospital events such as mortality and duration of stay. Another study showed that suicide attempts and completions could be predicted from five years of EHR demographic and clinical data (Simon et al., 2018).

4.3.4.2 Accuracy as a goal

Increasing machine learning's predictive accuracy can translate into patient safety- for example, the patient safety movement focusing on zero preventable deaths in hospitals by 2020 (https://patientsafetymovement.org/), is working on reducing the 1500 annual suicides in US hospitals. An effective machine learning classifier has been shown to contribute to this goal, predicting a patient's suicide attempts or completion (Simon et al., 2018).

Clinical questions that will benefit the most from machine learning as tools and techniques improve, have a common set of data characteristics contributing to better-performing models: the ability to collect an exponentially larger number of cases for model training, manually cleaned datasets, and training cases that clearly belong in a classification category of interest (Zhu et al., 2015).

4.3.4.3 Access to labeled datasets

Bipolar patients often require a change in therapy after weeks or months. With a complex, variable response phenotype, results suggest that no clear distribution of subjects response is currently available for bipolar disorders (Pisanu, Heilbronner & Squassina, 2018). In addition, the frequent changes to treatment that are required suggest a partial or incomplete response profile may be used as outcome labels on the training data. These techniques should account for changes in response state after measurement, requiring more data collection (e.g. by using streaming data past initial outcome measurements), in order to produce appropriate machine learning models predicting pharmacotherapy response. As Zhu et al. (2015) point out in their analysis of factors contributing to image classification performance, training examples that clearly correspond to a classification category contribute to model performance.

4.3.4.4 Objective label measurement

While new datasets are rapidly being made available, the response variables available in psychiatric data have been questioned, especially for binary predictions. Psychiatric treatment response data likely contain significant variability within subjects based on the heterogeneity of symptom profiles (Atkinson & Batterham, 2015), and a high measurement error of individual response to treatment (i.e. measuring the outcome variable with low precision), suggesting that carrying out classification tasks for these individuals would be inappropriate (Norbury, 2018). These sources of error become clear problems when attempting novel biomarker validation: reproducible results with high specificity and sensitivity are difficult to obtain, creating a low translation rate into the clinical environment (Drucker & Krapfenbauer, 2013).

4.4 KT & Private sector engagement & commercialization of software

4.4.1 Woebot

Woebot (https://woebot.io/) is a web-based cognitive-behavioural therapy (CBT; psychosocial therapy focused on reducing cognitive distortions) app designed to reduce therapy non-adherence through conversation abilities. RCT evidence demonstrated Woebot was able to demonstrate 27% less participant attrition than other online interventions for depression and anxiety (Fitzpatrick, Darcy & Vierhile, 2017). Woebot's effectiveness (e.g. 2.53 point drop in PHQ-9 score) was significant, and attributed to the bot's ability to express empathy. Other tools, such as DBT Coach for borderline personality disordered individuals, have demonstrated an average Beck Depression Inventory score reduction of 5.59 points after a similar treatment period (Rizvi et al., 2011).

4.4.2 Cognoa

Cognoa (https://www.cognoa.com/) is an automated childhood screening tool for Autism Spectrum Disorder (ASD) designed to offset the recent increase in wait times for ASD screening. It applies two machine learned classifiers to bin children into one of four risk categories for ASD (low, medium-non ASD, medium, and elevated): one classifier using parents' questionnaire responses, and another using an analyst's scores from a video recording of the child (Kanne, Carpenter & Warren, 2018). Cognoa performs at 71% accuracy (AUC 0.696), comparable to other scale-based ASD screening measures, but allows a single, smartphone-based test to be used for children under 6, while avoiding the multiple screening tests used current practice.

4.4.3 Babylon Health

Babylon (https://www.babylonhealth.com/) is a London-based AI company that developed a health app allowing patients to book virtual consultations with a GP (Iacobucci, 2017). Babylon

can access the EHR for each patient, look at treatment history, and assess disease symptoms prior to patients having a brief face-to-face conversation with a GP via smartphone video call. Preliminary work comparing Babylon's efficacy to human physicians found that it identified the condition with comparable precision and recall (Babylon F1 score of 57.1% vs physician F1 57.0% based on diagnosis of realistic clinical vignettes), and provided safer (97.0% vs 93.1%) triage advice, although the triage solutions were slightly less appropriate (0.5% more solutions fell outside the recommended range of triage options) than those recommended by human doctors (Razzaki et al., 2018).

4.4.4 Certification for use in mental health: standards of validity & efficacy

4.4.4.1 FDA

The FDA has released a Digital Health Innovation Action Plan (DHAP) to ensure the quality, safety, and efficacy of healthcare-focused digital technology (US FDA, 2018). This plan is focused on regulating digital health products that are assessed as a higher risk to patients, are made for specific conditions (as opposed to general wellness), and are not related to MDDS (Medical Device Data Systems; technologies that focus on medical data collection, storage, and movement) (US FDA, 2018b). The Software Precertification Pilot Program included in the DHAP is a streamlined process designed to accommodate medical software development. This program was created based on the FDA's assessment of their approach to moderate/high risk medical hardware development as lacking the agility and speed required to respond to medical software developers' needs. Part of this transition follows closely the evolution of medical devices from hardware-centric to a SaMD (Software as a Medical Device) approach. This program focuses on creating a responsive framework for FDA/developer communication, focusing on real-world software performance and KPI's (Key Performance Indicators), real-time consultations, and software quality management (Abram, 2017). Importantly for AI research, this

includes guidelines for how to carry out clinical trials, and how to assess AI system performance based on real-world measures (Jiang et al., 2017). In addition, the trials that assess AI are able to use an adaptive study design, enabling planned changes to the study design based on evidence collected during the trial (Graham, 2016).

4.4.4.2 Health Canada

Canada has been developing a similar plan (launching in 2020), emphasizing AI, telerobotics, and regulatory alignment with other HTA's (Health Technology Assessment organizations). These organizations, such as CADTH (Canadian Agency for Drugs and Technologies in Health), offer, "a comprehensive evaluation of the clinical effectiveness, cost-effectiveness, and the ethical, legal, and social implications of health technologies on patient health and the health care system" (CADTH, 2019). In addition, Health Canada's plan adds weight to considerations of cybersecurity, development of medical apps on mobile devices, and device interoperability. This plan also addresses care access and needs in rural and remote communities a priority, which will be an important consideration given the recent options to transition from face-to-face to mobile-based doctor appointments (Health Canada, 2017; Health Canada, 2018).

4.4.4.5 NHS (UK)

The UK's NHS Digital division has opted to harmonize medical device regulations with the European Commission's May 2017 regulations. These regulations begin treating software as a medical device, include risk to the patient as a measure in classifying new software, and like Canada, focus on cybersecurity, device interoperability, and mobile device platforms as delivery mechanisms (NHS, 2019).

4.5 Futurism & implications for ML in mental health

Understanding the details of how advanced machine learning tools (such as deep learning) work has been a major thrust of the artificial intelligence research community for years. The reasons behind it working have been explored with success, and improvements made to their expressibility, efficiency, and learnability (i.e. broader applications of neural networks that take less resources to run, and learn more quickly) (Lin, Tegmark & Rolnick, 2017). However, these tools have yet to integrate mental health to a significant degree: after assessing 309 health-related apps across Amazon's Alexa and Google's Assistant, only seven were focused on mental health (Chung et al., 2018).

4.5.1 Future vision: using AI as a full member of case consults ("Alexa in the room")

AI advances surrounding NLP (Natural Language Processing) allows review and analysis of EHR data. For example, ADEPt is data mining software that identifies a patient's adverse drug events from UK psychiatric EHRs with 83% accuracy (Iqbal et al., 2017). NLP data is difficult to use in the context of mental health, only recently becoming available in the form of neuropsychiatric clinical notes as part of the 2016 CEGS N-GRID Shared Task in Clinical Natural Language Programming (Filannino, Stubbs & Uzuner, 2017). These data have been used to create a predictor that is better than chance at using a patient's history of present illness to predict common mental conditions such as anxiety (Tran & Kavuluru, 2017).

4.5.2 The future looks data-driven: moving towards streaming data

Beyond formal EHR text, mining social media has been used with deep learning to extract psychiatric stressors for suicide, allowing for earlier suicidality detection and prevention (Du et al., 2018). On a larger stage, streaming data has the potential to change the practice of medicine. By combining a network of sensors (the "Internet of Things") into a proactive system, incipient diseases can be caught, personalized treatment created from tracking users' medical history, and

the financial and personnel burdens on healthcare reduced (Hassanalieragh et al., 2015). In mental health, this technology would work by training machine learning tools to correlate databases of mobile device sensor readings with users' psychiatric diagnoses, and using these trained tools to predict changes in psychiatric state based on current sensor readings (Hassanalieragh et al., 2015).

4.5.3 Exportability of AI solutions to manage off-site healthcare

Moving on to the global stage, smartphones are being applied to mental health during travel to other countries: one recent study tracked travellers' phones in Southeast Asia as a means of mapping spots where adverse health events were likely to occur (Farnham et al., 2018). This research built on a previous study by the same group, showing that daily tracking of health risks in Thailand using a mHealth app provided deeper data with less recall bias (errors in recollection) than post-trip questionnaires (Farnham et al., 2016). This is especially interesting from a mental health perspective, as a follow-up analysis showed that lethargy, anxiety, and irritability were all commonly reported symptoms during travel (at 80.0%, 34.7%, and 34.7%, respectively) (Farnham et al., 2017). The implications of this tracking could benefit mental health during travel, especially through the application of machine learning techniques that use phone sensors to capture data from travelers, and use it to create personalized behavioural recommendations to decrease mental health events (Sano et al., 2015; Bragazzi, Guglielmi, and Garbarino, 2019).

4.6 A note on search parameters

Throughout this review, we introduced papers across many domains of study. Some of the topics covered appeared to have few-to-no associated papers, until a shift in both search terms and search location was made. In the case of searching for ("covariate shift" + psychiat* + "machine learning"), for example, PubMed shows few results; the same is true for Google Scholar.

However, changing the term “psychiat*” to “antidepressant” reveals a literature on drug development and treatment prediction using psychiatry-specific datasets within Scholar only (e.g. Wen, Hassanpour & Greiner, 2016, using CO-MED data). We recommend that in the case of emerging interdomain topics, a sampling of common terms within the topic be used to find a more complete literature (e.g. using terms such as antipsychotic, anxiolytic, or drug discovery to better cover machine learning studies of psychiatry-related topics).

4.7 References

1. Abram AK. Fostering Medical Innovation: A Plan for Digital Health Devices; Software Precertification Pilot Program; 2017. govinfo.gov [Internet]. Available from: https://www.govinfo.gov/content/pkg/FR-2017-07-28/pdf/2017-15891.pdf
2. An Introduction to Health Technology Assessment | CADTH.ca [Internet]. CADTH. [cited 2019 Jun 28]. Available from: https://www.cadth.ca/introduction-health-technology-assessment
3. Atkinson G, Batterham AM. True and false interindividual differences in the physiological response to an intervention. Exp Physiol. 2015 Jun;100(6):577–88.
4. Bickel S, Brückner M, Scheffer T. Discriminative Learning Under Covariate Shift. J Mach Learn Res. 2009;10(Sep):2137–55.
5. Bragazzi NL, Guglielmi O, Garbarino AS. SleepOMICS: How Big Data Can Revolutionize Sleep Science. Int J Environ Res Public Health [Internet]. 2019 Jan 21;16(2). Available from: http://dx.doi.org/10.3390/ijerph16020291
6. Bull SA, Hunkeler EM, Lee JY, Rowland CR, Williamson TE, Schwab JR, et al. Discontinuing or switching selective serotonin-reuptake inhibitors. Ann Pharmacother. 2002 Apr;36(4):578–84.

7. Chen JH, Alagappan M, Goldstein MK, Asch SM, Altman RB. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. Int J Med Inform. 2017 Jun;102:71–9.

8. Chen JH, Asch SM. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations [Internet]. Vol. 376, New England Journal of Medicine. 2017. p. 2507–9. Available from: http://dx.doi.org/10.1056/nejmp1702071

9. Chung AE, Griffin AC, Selezneva D, Gotz D. Health and Fitness Apps for Hands-Free Voice-Activated Assistants: Content Analysis. JMIR Mhealth Uhealth. 2018 Sep 24;6(9):e174.

10. Clinical risk management standards - NHS Digital [Internet]. NHS Digital. [cited 2019 May 7]. Available from: https://digital.nhs.uk/services/solution-assurance/the-clinical-safety-team/clinical-risk-management-standards

11. Cummings JL, Morstorf T, Zhong K. Alzheimer's disease drug-development pipeline: few candidates, frequent failures [Internet]. Vol. 6, Alzheimer's Research & Therapy. 2014. p. 37. Available from: http://dx.doi.org/10.1186/alzrt269

12. Dataset Search [Internet]. [cited 2019 May 7]. Available from: https://toolbox.google.com/datasetsearch

13. Deo RC. Machine Learning in Medicine. Circulation. 2015 Nov 17;132(20):1920–30.

14. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ. 2016 May;47:20–33.

15. Doyle OM, Mehta MA, Brammer MJ. The role of machine learning in neuroimaging for drug discovery and development. Psychopharmacology . 2015 Nov;232(21-22):4179–89.

16. Drucker E, Krapfenbauer K. Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. EPMA J. 2013 Feb 25;4(1):7.

17. Drummond C, Holte RC. Cost curves: An improved method for visualizing classifier performance. Mach Learn. 2006 Oct 1;65(1):95–130.

18. Drummond C, Holte RC. Explicitly representing expected cost: An alternative to ROC representation. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2000. p. 198–207.

19. Drummond C, Holte RC. What ROC Curves Can't Do (and Cost Curves Can). In: ECAI [Internet]. Citeseer; 2004. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.240&rep=rep1&type=pdf

20. Du J, Zhang Y, Luo J, Jia Y, Wei Q, Tao C, et al. Extracting psychiatric stressors for suicide from social media using deep learning. BMC Med Inform Decis Mak. 2018 Jul 23;18(Suppl 2):43.

21. Esposito C, De Santis A, Tortora G, Chang H, Choo KR. Blockchain: A Panacea for Healthcare Cloud-Based Data Security and Privacy? IEEE Cloud Computing. 2018 Jan;5(1):31–7.

22. Farnham A, Blanke U, Stone E, Puhan MA, Hatz C. Travel medicine and mHealth technology: a study using smartphones to collect health data during travel. J Travel Med [Internet]. 2016 Jun;23(6). Available from: http://dx.doi.org/10.1093/jtm/taw056

23. Farnham A, Furrer R, Blanke U, Stone E, Hatz C, Puhan MA. The quantified self during travel: mapping health in a prospective cohort of travellers. J Travel Med [Internet]. 2017 Sep 1;24(5). Available from: http://dx.doi.org/10.1093/jtm/tax050

24. Farnham A, Röösli M, Blanke U, Stone E, Hatz C, Puhan MA. Streaming data from a smartphone application: A new approach to mapping health during travel. Travel Med Infect Dis. 2018 Jan;21:36–42.

25. Fernandez N, Copenhaver DJ, Vawdrey DK, Kotchoubey H, Stockwell MS. Smartphone Use Among Postpartum Women and Implications for Personal Health Record Utilization. Clin Pediatr . 2017 Apr;56(4):376–81.

26. Filannino M, Stubbs A, Uzuner Ö. Symptom severity prediction from neuropsychiatric clinical records: Overview of 2016 CEGS N-GRID shared tasks Track 2. J Biomed Inform. 2017 Nov;75S:S62–70.

27. Fitzpatrick KK, Darcy A, Vierhile M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial [Internet]. Vol. 4, JMIR Mental Health. 2017. p. e19. Available from: http://dx.doi.org/10.2196/mental.7785

28. Fortin J-P, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, et al. Harmonization of multi-site diffusion tensor imaging data. Neuroimage. 2017 Nov 1;161:149–70.

29. Fortney L. Blockchain, Explained [Internet]. Investopedia. 2019 [cited 2019 Jun 5]. Available from: https://www.investopedia.com/terms/b/blockchain.asp

30. Gartner Reprint [Internet]. [cited 2019 Feb 28]. Available from: https://www.gartner.com/doc/reprints?id=1-65WC0O1&ct=190128&st=sb

31. Graham J. Artificial Intelligence, Machine Learning, And The FDA. Forbes Magazine [Internet]. 2016 Aug 19 [cited 2019 May 7]; Available from: https://www.forbes.com/sites/theapothecary/2016/08/19/artificial-intelligence-machine-learning-and-the-fda/

32. Greenberg AJ, Haney D, Blake KD, Moser RP, Hesse BW. Differences in Access to and Use of Electronic Personal Health Information Between Rural and Urban Residents in the United States. J Rural Health. 2018 Feb;34 Suppl 1:s30–8.

33. Gressin S. The Equifax Data Breach: What to Do [Internet]. Consumer Information. 2017 [cited 2019 May 7]. Available from: https://www.consumer.ftc.gov/blog/2017/09/equifax-data-breach-what-do

34. Hajian S, Bonchi F, Castillo C. Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. p. 2125–6. (KDD '16).

35. Halevy A, Norvig P, Pereira F. The Unreasonable Effectiveness of Data [Internet]. Vol. 24, IEEE Intelligent Systems. 2009. p. 8–12. Available from: http://dx.doi.org/10.1109/mis.2009.36

36. Hassanalieragh M, Page A, Soyata T, Sharma G, Aktas M, Mateos G, et al. Health Monitoring and Management Using Internet-of-Things (IoT) Sensing with Cloud-Based Processing: Opportunities and Challenges. In: 2015 IEEE International Conference on Services Computing. 2015. p. 285–92.

37. Health Canada. Building better access to digital health technologies - Canada.ca [Internet]. 2017 [cited 2019 May 7]. Available from: https://www.canada.ca/en/health-canada/corporate/transparency/regulatory-transparency-and-openness/improving-review-drugs-devices/building-better-access-digital-health-technologies.html

38. Health Canada. Notice: Health Canada's Approach to Digital Health Technologies - Canada.ca [Internet]. 2018 [cited 2019 May 7]. Available from:

https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/activities/announcements/notice-digital-health-technologies.html

39. Heart T, Ben-Assuli O, Shabtai I. A review of PHR, EMR and EHR integration: A more personalized healthcare and public health policy. Health Policy and Technology. 2017 Mar 1;6(1):20–5.

40. Iacobucci G. London GP clinic sees big jump in patient registrations after Babylon app launch. BMJ. 2017 Dec 21;359:j5908.

41. Iniesta R, Malki K, Maier W, Rietschel M, Mors O, Hauser J, et al. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. J Psychiatr Res. 2016 Jul;78:94–102.

42. Iqbal E, Mallah R, Rhodes D, Wu H, Romero A, Chang N, et al. ADEPt, a semantically-enriched pipeline for extracting adverse drug events from free-text electronic health records [Internet]. Vol. 12, PLOS ONE. 2017. p. e0187121. Available from: http://dx.doi.org/10.1371/journal.pone.0187121

43. Jack. Copeland B. The Essential Turing. Clarendon Press; 2004. 620 p.

44. Jafari M, Safavi-Naini R, Sheppard NP. A rights management approach to protection of privacy in a cloud of electronic health records. In: Proceedings of the 11th annual ACM workshop on Digital rights management. ACM; 2011. p. 23–30.

45. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 May 2;13(6):395–405.

46. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017 Dec;2(4):230–43.

47. Kaelber DC, Jha AK, Johnston D, Middleton B, Bates DW. A research agenda for personal health records (PHRs). J Am Med Inform Assoc. 2008 Nov;15(6):729–36.

48. Kalia M, Costa E Silva J. Biomarkers of psychiatric diseases: current status and future prospects. Metabolism. 2015 Mar;64(3 Suppl 1):S11–5.

49. Kanne SM, Carpenter LA, Warren Z. Screening in toddlers and preschoolers at risk for autism spectrum disorder: Evaluating a novel mobile-health screening tool. Autism Res. 2018;11(7):1038–49.

50. Kaplan B. How Should Health Data Be Used?: Privacy, Secondary Use, and Big Data Sales. Camb Q Healthc Ethics. 2016 Apr;25(2):312–29.

51. Kostkova P, Brewer H, de Lusignan S, Fottrell E, Goldacre B, Hart G, et al. Who Owns the Data? Open Data for Healthcare [Internet]. Vol. 4, Frontiers in Public Health. 2016. Available from: http://dx.doi.org/10.3389/fpubh.2016.00007

52. Kostro D, Abdulkadir A, Durr A, Roos R, Leavitt BR, Johnson H, et al. Correction of inter-scanner and within-subject variance in structural MRI based automated diagnosing. Neuroimage. 2014 Sep;98:405–15.

53. Lee CH, Yoon H-J. Medical big data: promise and challenges. Kidney Res Clin Pract. 2017 Mar;36(1):3–11.

54. Lin HW, Tegmark M, Rolnick D. Why Does Deep and Cheap Learning Work So Well? J Stat Phys. 2017 Sep 1;168(6):1223–47.

55. Lozupone M, La Montagna M, D'Urso F, Daniele A, Greco A, Seripa D, et al. The Role of Biomarkers in Psychiatry. In: Guest PC, editor. Reviews on Biomarker Studies in Psychiatric and Neurodegenerative Disorders. Cham: Springer International Publishing; 2019. p. 135–62.

56. Malki K, Koritskaya E, Harris F, Bryson K, Herbster M, Tosto MG. Epigenetic differences in monozygotic twins discordant for major depressive disorder. Transl Psychiatry. 2016 Jun 14;6(6):e839.

57. McLellan RA, Oscarson M, Seidegård J, Evans DA, Ingelman-Sundberg M. Frequent occurrence of CYP2D6 gene duplication in Saudi Arabians. Pharmacogenetics. 1997 Jun;7(3):187–91.

58. Minsky, Marvin (1967), Computation: Finite and Infinite Machines, Englewood Cliffs, N.J.: Prentice-Hall.

59. Moravec HP. The Stanford Cart and the CMU Rover. Proc IEEE. 1983 Jul;71(7):872–84.

60. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013 Apr 3;309(13):1351–2.

61. Norbury A. Response heterogeneity: Challenges for personalised medicine and big data approaches in psychiatry and chronic pain. F1000Res [Internet]. 2018 [cited 2019 May 7];7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5820606/

62. O'Doherty KC, Christofides E, Yen J, Bentzen HB, Burke W, Hallowell N, et al. If you build it, they will come: unintended future uses of organised health data collections [Internet]. Vol. 17, BMC Medical Ethics. 2016. Available from: http://dx.doi.org/10.1186/s12910-016-0137-x

63. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016 Sep 29;375(13):1216–9.

64. Piatetsky G. Gainers, Losers, and Trends in Gartner 2019 Magic Quadrant for Data Science and Machine Learning Platforms [Internet]. 2019 [cited 2019 Feb 28]. Available

from: https://www.kdnuggets.com/2019/02/gartner-2019-mq-data-science-machine-learning-changes.html

65. Pisanu C, Heilbronner U, Squassina A. The Role of Pharmacogenomics in Bipolar Disorder: Moving Towards Precision Medicine [Internet]. Vol. 22, Molecular Diagnosis & Therapy. 2018. p. 409–20. Available from: http://dx.doi.org/10.1007/s40291-018-0335-y

66. Quiñonero-Candela J. Dataset Shift in Machine Learning. MIT Press; 2009. 229 p.

67. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine. 2018 May 8;1(1):18.

68. Razzaki S, Baker A, Perov Y, Middleton K, Baxter J, Mullarkey D, et al. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis [Internet]. arXiv [cs.AI]. 2018. Available from: http://arxiv.org/abs/1806.10698

69. Rizvi SL, Dimeff LA, Skutch J, Carroll D, Linehan MM. A pilot study of the DBT coach: an interactive mobile phone application for individuals with borderline personality disorder and substance use disorder. Behav Ther. 2011 Dec;42(4):589–600.

70. Roehrs A, da Costa CA, da Rosa Righi R. OmniPHR: A distributed architecture model to integrate personal health records. J Biomed Inform. 2017 Jul;71:70–81.

71. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958 Nov;65(6):386–408.

72. Sano A, Phillips AJ, Yu AZ, McHill AW, Taylor S, Jaques N, et al. Recognizing Academic Performance, Sleep Quality, Stress Level, and Mental Health using Personality Traits, Wearable Sensors and Mobile Phones. Int Conf Wearable Implant Body Sens Netw [Internet]. 2015 Jun;2015. Available from: http://dx.doi.org/10.1109/BSN.2015.7299420

73. Sejnowski TJ, Rosenberg CR. Parallel networks that learn to pronounce English text. Complex Systems. 1987;1(1):145–68.

74. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform. 2018 Sep;22(5):1589–604.

75. Shortliffe EH, Buchanan BG. A Model of Inexact Reasoning in Medicine. Math Biosci. 1975 Apr 1;23(3-4):351–79.

76. Simon GE, Johnson E, Lawrence JM, Rossom RC, Ahmedani B, Lynch FL, et al. Predicting Suicide Attempts and Suicide Deaths Following Outpatient Visits Using Electronic Health Records. Am J Psychiatry. 2018 Oct 1;175(10):951–60.

77. Sittig DF, Singh H. Legal, ethical, and financial dilemmas in electronic health record adoption and use. Pediatrics. 2011 Apr;127(4):e1042–7.

78. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, et al. Grand challenges in clinical decision support. J Biomed Inform. 2008 Apr;41(2):387–92.

79. Soni D. Dealing with Imbalanced Classes in Machine Learning [Internet]. Towards Data Science. Towards Data Science; 2018 [cited 2019 May 7]. Available from: https://towardsdatascience.com/dealing-with-imbalanced-classes-in-machine-learning-d43d6fa19d2

80. Spencer K, Sanders C, Whitley EA, Lund D, Kaye J, Dixon WG. Patient Perspectives on Sharing Anonymized Personal Health Data Using a Digital System for Dynamic Consent and Research Feedback: A Qualitative Study. J Med Internet Res. 2016 Apr 15;18(4):e66.

81. Terry AL, Stewart M, Fortin M, Wong ST, Kennedy M, Burge F, et al. Gaps in primary healthcare electronic medical record research and knowledge: findings of a pan-Canadian study. Healthc Policy. 2014;10(1):46–59.

82. Torous JB, Chan SR, Gipson SY-MT, Kim JW, Nguyen T-Q, Luo J, et al. A Hierarchical Framework for Evaluation and Informed Decision Making Regarding Smartphone Apps for Clinical Care. Psychiatr Serv. 2018 May 1;69(5):498–500.

83. Tran T, Kavuluru R. Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks. J Biomed Inform. 2017 Nov 1;75:S138–48.

84. U.S. Food and Drug Administration. Digital Health Innovation Action Plan. 2018.

85. U.S. Food and Drug Administration. Digital health software precertification (Pre-Cert) program. Available at:)(Accessed May 27, 2018) https://www fda gov/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/default htm View in Article. 2018b.

86. Vassos E, Agerbo E, Mors O, Pedersen CB. Urban–rural differences in incidence rates of psychiatric disorders in Denmark. Br J Psychiatry. 2016 May;208(5):435–40.

87. Wen J, Hassanpour N, Greiner R. Weighted Gaussian Process for Estimating Treatment Effect. 30th Conference on Neural Information Processing Systems [Internet]. 2016; Available from: https://pdfs.semanticscholar.org/9706/6ad9f744213f54e1afa04c991055b1345f03.pdf

88. Wikipedia contributors. AI winter [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Jul 4]. Available from: https://en.wikipedia.org/w/index.php?title=AI_winter&oldid=903461097

89. Wikipedia contributors. Arthur Samuel [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Jul 4]. Available from: https://en.wikipedia.org/w/index.php?title=Arthur_Samuel&oldid=887136586

90. Wikipedia contributors. Blockchain [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Jun 5]. Available from: https://en.wikipedia.org/w/index.php?title=Blockchain&oldid=900245799

91. Wikipedia contributors. Grace Hopper [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Jul 4]. Available from: https://en.wikipedia.org/w/index.php?title=Grace_Hopper&oldid=904577918

92. Zapata BC, Niñirola AH, Fernández-Alemán JL, Toval A. Assessing the privacy policies in mobile personal health records. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2014. p. 4956–9.

93. Zhu X, Vondrick C, Fowlkes C, Ramanan D. Do We Need More Training Data? [Internet]. arXiv [cs.CV]. 2015. Available from: http://arxiv.org/abs/1503.01508

Connection Between Chapters 4-5

The impact on psychiatric practice caused by the deployment of machine learning-based tools has not yet been assessed. However, with such a disruptive and broadly applicable technology on the horizon, changes resulting from the deployment of these tools is now an area of growing interest. The next chapter builds on the description of ML in personalized medicine and psychiatry in Chapter 3, and establishes a viewpoint for the future: how deploying machine learning in the clinic will change prescribing practice in psychiatry.

Choosing which antidepressant to prescribe is currently based on a clinician's experience with the choices available, and the tolerability of that choice to the patient. Statistical support to inform those choices is available from clinical trial data, including information such as medication efficacy, safety, and side-effect tolerability. However, these comparisons are made at a population level, without accounting for individual differences in medication response, and do not include measures of outcome certainty (analogous to predictive accuracy). At best, these data support prescription of the medication most likely to be effective for patients in general – but does not allow individualized prescriptions, determining if a specific patient will benefit. If machine learning models are put into clinical use, they stand to disrupt this treatment model by providing clinicians with patient-specific outcome predictions with a known failure rate. The following chapter discusses possible effects of this disruption at a patient, physician, and pharmaceutical company R&D level.

Chapter 5. From Efficacy to Accuracy: How Machine Learning Will Change Prescribing Practice in Depression

James R.A. Benoit, MA[1*], Russell Greiner, PhD[2], Serdar M. Dursun, MD[1]

[1] Department of Psychiatry, University of Alberta, 1E1 Walter Mackenzie Health Sciences Centre, 8440 112 St NW, Edmonton, Alberta, Canada, T6G 2B7

[2] Department of Computing Science, University of Alberta, 2-32 Athabasca Hall, Edmonton, Alberta, Canada, T6G 2E8

[*] Corresponding author

Clinical trials for antidepressants generate three useful endpoints for clinical decision making: statistics of effectiveness, safety, and tolerability. Effectiveness measures an antidepressant's function in real-world settings, while safety involves pharmacovigilance identifying the antidepressant's adverse effects, and tolerability examines how well those adverse effects are withstood by patients. These measures are fundamental for the development of treatment recommendations such as the American Psychiatric Association's guidelines for treating major depressive disorder (MDD) (Gelenberg et al., 2010).

Prescribing practice for depression is a recent focus of machine learning research, which aims to predict antidepressant treatment outcomes at baseline from clinical data (Chekroud et al., 2016). The motivation behind this research is to avoid therapeutic delays by accurately prescribing an effective, safe, and well-tolerated antidepressant on the first try. This approach can use as input all the data outputs of a clinical trial: primarily patient demographics, lab tests, and psychiatric scale data. These data have two dimensions: the number of patient cases, and number of clinical measurements (in machine learning parlance, "features") for each patient. As part of creating a machine learning model, clinical measurements are assessed agnostically (i.e. without bias based on previous knowledge of how those measures are expected to vary between patients) for whether their inclusion in the treatment outcome prediction model will be beneficial to the model's predictive accuracy. The quality of feature assessment as well as the accuracy of the final model is improved as sample size increases (Patel, Khalaf & Aizenstein, 2016); as the quality of data fed into a machine learning algorithm improves (i.e. it has few missing datapoints or erroneous entries), the quality of model produced by the algorithm improves.

Therefore, machine learning models predicting treatment outcome in patients with MDD will be improved as the volume and quality of data available to them increases. As these models move

from the lab bench toward deployment in real-world healthcare settings (e.g. hospitals, clinics, and telehealth) there will be a range of models available as clinical decision support tools to assist psychiatrists' prescription decisions. These models will assess the likelihood that their patient's depression will respond to a particular antidepressant. Based on the amount and quality of data that went into each model, treatment outcome prediction models for some antidepressants will be more accurate than other models. This begs the question: if one antidepressant can be prescribed more accurately than another, do physicians have a duty to their patients to first consider prescription of antidepressants with the most accurate models predicting treatment outcome?

Currently, prescription of antidepressants is currently a trial-and-error process: clinical trial meta-analyses for newer medications found that 37% of patients do not respond and 53% do not remit following 6-12 weeks of treatment (Gartlehner et al., 2011). There is no way of accurately predicting treatment outcome for an individual, aside from relying on a clinician's experience with a particular antidepressant. However, accurate models predicting treatment outcome are on the horizon: The machine learned model described in Chekroud et al. (2016) predicted treatment response to citalopram with 64.6% accuracy, and Iniesta et al. (2016)'s machine learned model predicted remission following escitalopram or nortriptyline treatment with ROC AUC's of 0.75 and 0.72, respectively (Iniesta et al., 2016). With these early successes, it is clear that using machine learning-based models of treatment response prediction is a viable pathway towards producing better treatment outcomes for patients. When these models are deployed in the real world, physicians will have a duty to use them in prescribing antidepressants: they reduce the risk of harm to patients and provide a measurable benefit to the MDD patient population.

If we take it as true that (1) these treatment outcome prediction models will eventually be used in a clinical setting, (2) the models with the most patient cases informing them will outperform other models, and (3) clinicians will continue to carefully administer their duty of care, it follows that antidepressants with the lowest expected cost (where cost is defined as detriment to the patient) models available for predicting their treatment outcomes should be considered for use before antidepressants with more costly models underlying their treatment outcomes.

In an ideal setting, where multiple models of treatment outcome prediction are available, all models would be considered before prescribing an antidepressant. The reality in the clinic is that time is a limited resource, and information gathering for the features used by these models (e.g. lab tests, physician-administered psychiatric scales) should be made as cost-effective as possible. There is no guarantee that different models of treatment outcome prediction will need the same information from patients: looking at the field as a whole, it is apparent that different models will invariably differ in the information they need to make accurate predictions (Shatte, Hutchinson & Teague, 2019). Therefore, collecting clinical features from patients relevant to the most accurate models predicting antidepressant treatment outcomes will be prioritized.

Prioritizing the probability that a particular antidepressant will work on a particular patient (i.e. the prescribed drug with the lowest expected cost to the patient) over the group efficacy of an antidepressant will create a new objective in which pharmaceutical research holding the best evidence for accurately predicting treatment outcome of their antidepressant, will see it being the first considered for prescription. If the most accurately prescribed drug is considered first, it lays the groundwork for a new competition between pharmaceutical companies: a race towards accurate outcome prediction tools, fueled by data. See Figure 5.1 for an example use case.

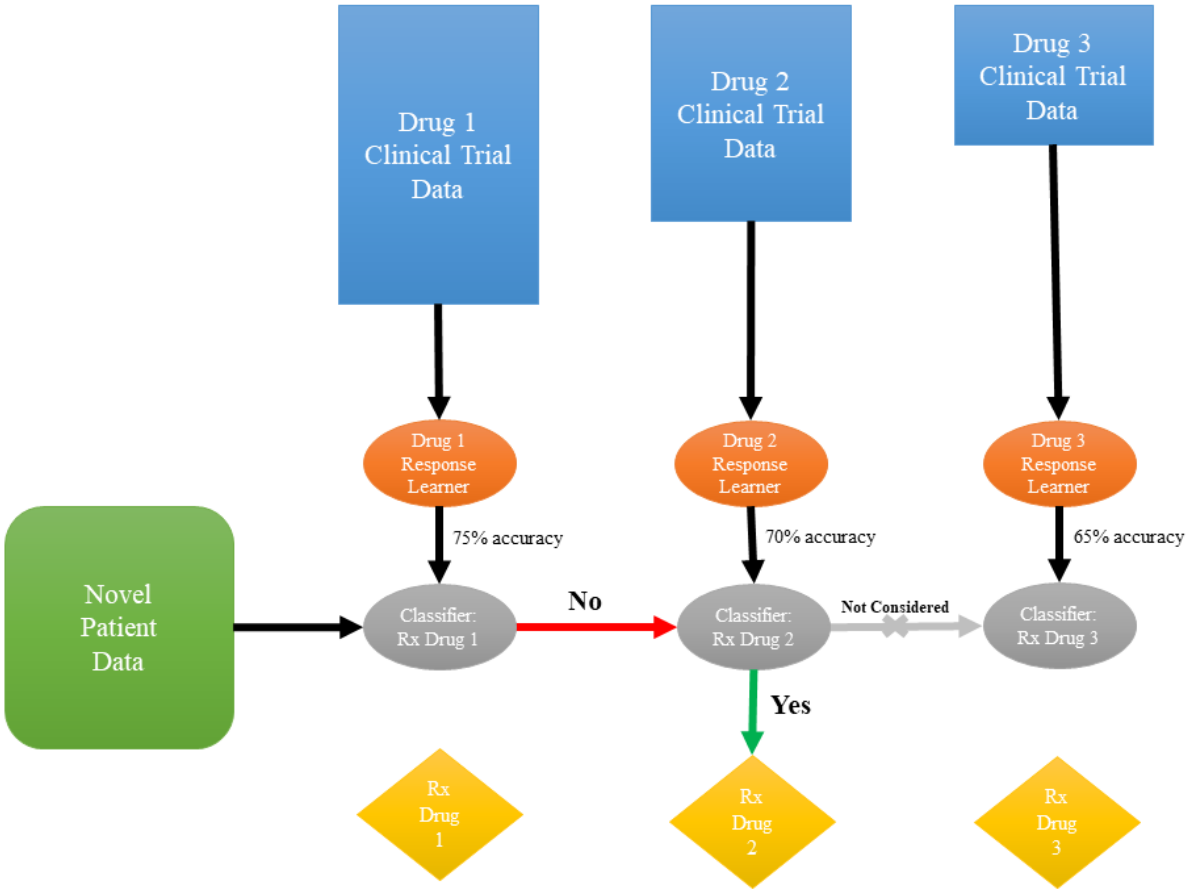

Figure 5.1. Use case example for classification accuracy-based prescription system

Data from an established antidepressant can generate a treatment outcome prediction model that is superior to the model for other antidepressants and will therefore be widely considered for prescription first. Even if the model only suggests a drug be prescribed in a quarter of assessed patients, that antidepressant will still maintain a 25% share of the patient demographic among all available antidepressants. Given the 21 widely available antidepressants currently available to patients (Cipriani et al., 2018), a 25% capture of potential patients will provide a significant incentive in pharmaceutical development to compete to produce the most accurate model predicting treatment outcomes (the 75% of patients not prescribed the first considered

antidepressant would then be tested on additional antidepressant treatment outcome prediction models until one is found that supports the patient's response to that treatment).

Patients stand to benefit from this scenario: patients' needs will be better served because they will be prescribed effective antidepressant medications with higher accuracy. Clinicians will have well-validated tools produced by a competition for predictive accuracy. Pharmaceutical R&D will compete to collect as much data as possible for predicting antidepressant treatment outcomes, which will lead to novel research being produced on depression diagnosis and treatment. At the same time, these benefits will not come without drawbacks: in choosing among antidepressants to be considered for prescription first, new medications will have difficulty obtaining the volume of evidence necessary. As a result, a few antidepressants may come to dominate the market, and possibly stagnate further antidepressant development. However, if models of treatment outcome converge on a set of common clinical features, concurrent treatment outcome assessment of multiple antidepressants (including new medications) may become possible.

Another solution to the issue of having multiple models of treatment outcome involves active classifiers. These are systems specialized in ranking features in order of how much they contribute to reducing the system's error. Active classifiers can also stop mid-classification to ask the human (referred to as the "oracle") for pieces of missing information. These are a natural fit with systems such as the one suggested above, which promote democratized access to healthcare, because they are capable of adapting to healthcare systems that are resource, equipment, and personnel-scarce, with minimal patient impact (Greiner et al., 2002; Settles 2009).

With wearables and other devices in the Internet of Things (Wikipedia contributors, 2019) producing personalized health data, patient electronic health records becoming more widely available to researchers, and clinical trial data becoming more accessible, a favourable environment has been created for machine learning innovation surrounding antidepressant treatment outcome prediction. With the advent of treatment outcome prediction models being deployed in clinical settings, the psychiatric prescription model will begin shifting away from overall antidepressant efficacy towards prescription accuracy, changing antidepressant prescribing practice for major depressive disorder from trial-and-error to machine learning-informed.

5.1 References

1. Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*. 2016;3(3):243-250.
2. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet*. 2018;391(10128):1357-1366.
3. Gartlehner G, Hansen RA, Morgan LC, et al. Comparative benefits and harms of second-generation antidepressants for treating major depressive disorder: an updated meta-analysis. *Ann Intern Med*. 2011;155(11):772-785.
4. Gelenberg AJ, Freeman MP, Markowitz JC, et al. Practice guideline for the treatment of patients with major depressive disorder third edition. *Am J Psychiatry*. 2010;167(10):1.
5. Greiner R, Grove AJ, Roth D. Learning cost-sensitive active classifiers. Artif Intell. 2002;139(2):137–74.

6. Iniesta R, Malki K, Maier W, et al. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *J Psychiatr Res*. 2016;78:94-102.

7. Patel MJ, Khalaf A, Aizenstein HJ. Studying depression using imaging and machine learning methods. *Neuroimage Clin*. 2016;10:115-123.

8. Settles B. Active learning literature survey [Internet]. University of Wisconsin-Madison Department of Computer Sciences; 2009. Available from: https://minds.wisconsin.edu/handle/1793/60660

9. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. *Psychol Med*. February 2019:1-23.

10. Wikipedia contributors. Internet of things [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Jun 5]. Available from: https://en.wikipedia.org/w/index.php?title=Internet_of_things&oldid=900133956

Chapter 6. Conclusion

The goal of this thesis was to contribute to computational psychiatry and data-driven medicine by demonstrating the viability of machine learning for clinical use, focusing on predicting treatment outcomes in MDD. To support this goal, I demonstrated through two original research articles that predicting symptom remission and treatment response are possible above chance using machine learning software, used a conceptual review to provide an overview of what other machine learning -based solutions are being developed in the current healthcare environment, and extended how these solutions might affect clinical practice in the future through a viewpoint article. In addition, I used Python to develop and publicly release a machine learning pipeline for predicting treatment response that can be applied to any clinical trial data (see Appendix 1).

Machine learning has the potential to advance personalized, data-driven psychiatric care. It is approaching deployment in a clinical environment for predicting patient response to antidepressant treatment. This is supported by the development of FDA and IMDRF regulatory frameworks for adaptive algorithms, the increasing number of companies devoted to ML-based health solutions, and the exponential increase in research publications in the field of data-driven medicine over the last 10 years. The deployment of ML-based software in psychiatry stands to benefit all strata of the healthcare environment, and should be considered from a number of different perspectives, including:

1. Patients and patient families
2. Clinicians, specifically psychiatrists
3. Healthcare insurers
4. Provincial health authorities

5. Federal regulatory bodies
6. Clinically-oriented ML research

One of the most important considerations for ML-based tools is how they will be perceived and trusted by patients. Inherently, tools that offer objective predictions for how the world works, have been shown to make more accurate predictions of treatment outcome than physicians, and have a known error rate. However, the clinical reality of the importance of a physician-patient relationship may be better reflected by Balint, "...by far the most frequently used drug in general practice was *the doctor himself*." (Balint, 1955) Data from the uptake and acceptance of these tools after broad integration into practice will be a welcome addition to our knowledge of patient-algorithm interactions. Interestingly, machine learning-based solutions in Canada may stand to benefit three marginalized patient populations the most. Rural, arctic, and indigenous patients have significant barriers of access to psychiatric care: geographical, cultural, and socio-economic (Marrone, 2007). Enabling them to receive the same quality recommendations for treatment as patients living in urban centers will improve their agency as patients.

From a psychiatrist's perspective, the addition of data-driven tools for forecasting treatment outcomes may have wide-ranging effects on prescribing practice: as discussed in Chapter 4, these tools may impact which drugs are considered first for prescription. Physicians will be asked to place their trust in a new technology, and the incorporation of safety, efficacy, and risk management requirements for tools in the proposed FDA SaME regulatory framework (FDA, 2019), suggest that the elements contributing to that trust will be a focus of the first wave of adaptive (as opposed to deterministic) algorithms for healthcare.

Building on the theme of trust, one element not yet considered is how tool errors will contribute to the public perception of integrating ML into clinical practice. As discussed in chapter 3, a hard lesson learned from the first AI winter in the 1970's will be to mitigate drops in public confidence stemming from over-promising and under-delivering on ML tool performance (Wikipedia contributors, 2019). This is especially important when interacting with vulnerable populations such as depressed patients, where visible failures to treat with novel technology have more potential to produce public outcry than established treatments.

One of the most marked improvements yet to be implemented in machine learning is determining which features will contribute the most to an algorithm's predictive accuracy. While there is ample research into face and construct validity of psychiatric questionnaires, these tools do not take into account how a machine learning algorithm may interpret data (e.g. there has been no attempt to determine how re-wording questions affects predictive accuracy). To accomplish this, clinical, lab, and demographic data must be made more available to researchers (Ross, Lehman & Gross, 2012). Initiatives such as the Clinical Study Data Request (ClinicalStudyDataRequest.com, 2019), are attempting to remedy this lack of access.

The work in this thesis should be extended in several ways. First, clinical trial data from more antidepressants should be evaluated using the trained classifier $C_{Concure}$ produced in Chapter 2 to determine how generalizable this tool is across medications, and the stability of features chosen by the model across medications. Second, more finely grained definitions of patient populations (e.g. divided geographically, by medication response time, by depression severity, or divided using a computational technique such as k-means clustering) should be compared, to determine the relative effectiveness of using one classifier trained on all patients' data compared to a

plethora of classifiers, each trained on the data from one group of patients. This will allow assessment of dataset size requirements for effective classifier training. Third, moving from running a shell-based machine learning program to an mHealth (mobile Health) app would greatly decrease the learning curve and accessibility of this software for non-domain experts. Since this type of tool has not yet been deployed in a clinical setting, there is no data for how it will be adopted by clinicians, its intended vs. real-world use cases, or changes that will need to be made in order to make it a viable addition to the treatment decision-making process. Fourth, future machine learning tools produced should integrate and test new learners such as those found in the PyTorch and TensorFlow libraries.

Aside from classifier-related improvements, adding more to features' predictive value could be accomplished through assessing the nature of data being collected in clinical trials. While data is collected based on well-established, validated measures of patient demographics, lab tests, and psychiatric scales, the validation process supporting these measures is based on techniques drawn from association studies; for example, the HAM-D uses questions that have been shown, on average, to elucidate specific dimensions of a patient's depression (Rush et al., 2006; Bobo et al., 2016). It is unknown whether these narrow categories would serve a classifier as well as asking more broadly worded or differently phrased questions at the time of data collection.

In order to generate classifiers that are trained on the same distribution of data they are expected to draw on for predictions, determining the composition of real-world patient populations will allow for classifier creation that reflects a clinical environment. While clinical trials provide a large, organized base of data on which to train classifiers, there is a significant difference between patients who enter a clinical trial (e.g. meeting certain exclusion criteria such as a lack

of previous antidepressant use), and those entering a clinic for treatment. The cost of misclassifying real-world patients should also be assigned a meaningful value, in order to better determine which classifier would best serve the patient population. In Chapter 3, we examined the benefit of using cost curves to determine when to use each classifier, but these curves cannot be used in machine learning tool development until a population's misclassification costs are known.

While there are many areas of machine learning that merit further inquiry, it shows enormous promise for contributing to treatment outcome prediction in psychiatry. It is exciting to consider how data-driven approaches will benefit future patients, creating a new frontier to explore improvements in mental health, seek out new ways to alleviate strife in those suffering from depression, and boldly go where medicine has not gone before.

6.1 References

1. Balint M. The doctor, his patient, and the illness. Lancet. 1955 Apr 2;268(6866):683–8.
2. Bobo WV, Angleró GC, Jenkins G, Hall-Flavin DK, Weinshilboum R, Biernacka JM. Validation of the 17-item Hamilton Depression Rating Scale definition of response for adults with major depressive disorder using equipercentile linking to Clinical Global Impression scale ratings: analysis of Pharmacogenomic Research Network Antidepressant Medication Pharmacogenomic Study (PGRN-AMPS) data. Hum Psychopharmacol. 2016 May;31(3):185–92.
3. ClinicalStudyDataRequest.com [Internet]. [cited 2019 Apr 29]. Available from: https://www.clinicalstudydatarequest.com/Default.aspx

4. Marrone S. Understanding barriers to health care: a review of disparities in health care services among indigenous populations. Int J Circumpolar Health. 2007 Jun;66(3):188–98.
5. Ross JS, Lehman R, Gross CP. The importance of clinical trial data sharing: toward more open science. Circ Cardiovasc Qual Outcomes. 2012 Mar 1;5(2):238–40.
6. Rush AJ, Kraemer HC, Sackeim HA, Fava M, Trivedi MH, Frank E, Ninan PT, Thase ME, Gelenberg AJ, Kupfer DJ, Regier DA, Rosenbaum JF, Ray O, Schatzberg AF. Report by the ACNP Task Force on response and remission in major depressive disorder. Neuropsychopharmacology. 2006;31(9):1841–1853.
7. US FDA Artificial Intelligence and Machine Learning Discussion Paper. Available from: https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/SoftwareasaMedicalDevice/UCM635052.pdf
8. Wikipedia contributors. AI winter [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Apr 29]. Available from: https://en.wikipedia.org/w/index.php?title=AI_winter&oldid=894747522

Bibliography

1. Abram AK. Fostering Medical Innovation: A Plan for Digital Health Devices; Software Precertification Pilot Program; 2017. govinfo.gov [Internet]. Available from: https://www.govinfo.gov/content/pkg/FR-2017-07-28/pdf/2017-15891.pdf
2. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR) [Internet]. 2000. Available from: http://dx.doi.org/10.1176/appi.books.9780890423349
3. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Pub; 2013. 991 p.
4. American Psychiatric Organization (APA). DSM History [Internet]. [cited 2019 Jun 10]. Available from: https://www.psychiatry.org/psychiatrists/practice/dsm/history-of-the-dsm
5. An Introduction to Health Technology Assessment | CADTH.ca [Internet]. CADTH. [cited 2019 Jun 28]. Available from: https://www.cadth.ca/introduction-health-technology-assessment
6. Atkinson G, Batterham AM. True and false interindividual differences in the physiological response to an intervention. Exp Physiol. 2015 Jun;100(6):577–88.
7. Bados A, Balaguer G, Saldaña C. The efficacy of cognitive-behavioral therapy and the problem of drop-out. J Clin Psychol. 2007 Jun;63(6):585–92.
8. Bagby RM, Ryder AG, Schuller DR, Marshall MB. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? Am J Psychiatry [Internet]. 2004 Dec;161(12):2163–77. Available from: http://dx.doi.org/10.1176/appi.ajp.161.12.2163
9. Balint M. The doctor, his patient, and the illness. Lancet. 1955 Apr 2;268(6866):683–8.

10. Bickel S, Brückner M, Scheffer T. Discriminative Learning Under Covariate Shift. J Mach Learn Res. 2009;10(Sep):2137–55.

11. Bobo WV, Angleró GC, Jenkins G, Hall-Flavin DK, Weinshilboum R, Biernacka JM. Validation of the 17-item Hamilton Depression Rating Scale definition of response for adults with major depressive disorder using equipercentile linking to Clinical Global Impression scale ratings: analysis of Pharmacogenomic Research Network Antidepressant Medication Pharmacogenomic Study (PGRN-AMPS) data. Hum Psychopharmacol. 2016 May;31(3):185–92.

12. Bragazzi NL, Guglielmi O, Garbarino AS. SleepOMICS: How Big Data Can Revolutionize Sleep Science. Int J Environ Res Public Health [Internet]. 2019 Jan 21;16(2). Available from: http://dx.doi.org/10.3390/ijerph16020291

13. Brown G, Pocock A, Zhao M-J, Luján M. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. J Mach Learn Res [Internet]. 2012 [cited 2019 Feb 7];13(Jan):27–66. Available from: http://www.jmlr.org/papers/volume13/brown12a/brown12a.pdf

14. Brown MRG, Sidhu GS, Greiner R, Asgarian N, Bastani M, Silverstone PH, et al. ADHD-200 Global Competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. Front Syst Neurosci [Internet]. 2012 Sep 28;6:69. Available from: http://dx.doi.org/10.3389/fnsys.2012.00069

15. Brown TA, Di Nardo PA, Lehman CL, Campbell LA. Reliability of DSM-IV anxiety and mood disorders: implications for the classification of emotional disorders. J Abnorm Psychol. 2001 Feb;110(1):49–58.

16. Bull SA, Hunkeler EM, Lee JY, Rowland CR, Williamson TE, Schwab JR, et al. Discontinuing or switching selective serotonin-reuptake inhibitors. Ann Pharmacother. 2002 Apr;36(4):578–84.

17. Burton R. The Anatomy of Melancholy: What it Is, with All the Kinds, Causes, Symptoms, Prognostics and Several Cures of it. John C. Nimmo; 1886. 558 p.

18. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry. 2016 Mar;3(3):243–50.

19. Chen JH, Alagappan M, Goldstein MK, Asch SM, Altman RB. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. Int J Med Inform. 2017 Jun;102:71–9.

20. Chen JH, Asch SM. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations [Internet]. Vol. 376, New England Journal of Medicine. 2017. p. 2507–9. Available from: http://dx.doi.org/10.1056/nejmp1702071

21. Chung AE, Griffin AC, Selezneva D, Gotz D. Health and Fitness Apps for Hands-Free Voice-Activated Assistants: Content Analysis. JMIR Mhealth Uhealth. 2018 Sep 24;6(9):e174.

22. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. Lancet. 2018;391(10128):1357-1366.

23. Clinical risk management standards - NHS Digital [Internet]. NHS Digital. [cited 2019 May 7]. Available from: https://digital.nhs.uk/services/solution-assurance/the-clinical-safety-team/clinical-risk-management-standards

24. ClinicalStudyDataRequest.com [Internet]. [cited 2019 Apr 29]. Available from: https://www.clinicalstudydatarequest.com/Default.aspx

25. Cummings JL, Morstorf T, Zhong K. Alzheimer's disease drug-development pipeline: few candidates, frequent failures [Internet]. Vol. 6, Alzheimer's Research & Therapy. 2014. p. 37. Available from: http://dx.doi.org/10.1186/alzrt269

26. Dataset Search [Internet]. [cited 2019 May 7]. Available from: https://toolbox.google.com/datasetsearch

27. Deecher DC, Beyer CE, Johnston G, Bray J, Shah S, Abou-Gharbia M, et al. Desvenlafaxine succinate: A new serotonin and norepinephrine reuptake inhibitor. J Pharmacol Exp Ther [Internet]. 2006 Aug;318(2):657–65. Available from: http://dx.doi.org/10.1124/jpet.106.103382

28. Deo RC. Machine Learning in Medicine. Circulation. 2015 Nov 17;132(20):1920–30.

29. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ. 2016 May;47:20–33.

30. Doyle OM, Mehta MA, Brammer MJ. The role of machine learning in neuroimaging for drug discovery and development. Psychopharmacology . 2015 Nov;232(21-22):4179–89.

31. Drucker E, Krapfenbauer K. Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. EPMA J. 2013 Feb 25;4(1):7.

32. Drummond C, Holte RC. Cost curves: An improved method for visualizing classifier performance. Mach Learn. 2006 Oct 1;65(1):95–130.

33. Drummond C, Holte RC. Explicitly representing expected cost: An alternative to ROC representation. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2000. p. 198–207.

34. Drummond C, Holte RC. What ROC Curves Can't Do (and Cost Curves Can). In: ECAI [Internet]. Citeseer; 2004. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.240&rep=rep1&type=pdf

35. Du J, Zhang Y, Luo J, Jia Y, Wei Q, Tao C, et al. Extracting psychiatric stressors for suicide from social media using deep learning. BMC Med Inform Decis Mak. 2018 Jul 23;18(Suppl 2):43.

36. Ersoy S, Engin VS. Risk factors for polypharmacy in older adults in a primary care setting: a cross-sectional study. Clin Interv Aging [Internet]. 2018 Oct 15;13:2003–11. Available from: http://dx.doi.org/10.2147/CIA.S176329

37. Esposito C, De Santis A, Tortora G, Chang H, Choo KR. Blockchain: A Panacea for Healthcare Cloud-Based Data Security and Privacy? IEEE Cloud Computing. 2018 Jan;5(1):31–7.

38. Farnham A, Blanke U, Stone E, Puhan MA, Hatz C. Travel medicine and mHealth technology: a study using smartphones to collect health data during travel. J Travel Med [Internet]. 2016 Jun;23(6). Available from: http://dx.doi.org/10.1093/jtm/taw056

39. Farnham A, Furrer R, Blanke U, Stone E, Hatz C, Puhan MA. The quantified self during travel: mapping health in a prospective cohort of travellers. J Travel Med [Internet]. 2017 Sep 1;24(5). Available from: http://dx.doi.org/10.1093/jtm/tax050

40. Farnham A, Röösli M, Blanke U, Stone E, Hatz C, Puhan MA. Streaming data from a smartphone application: A new approach to mapping health during travel. Travel Med Infect Dis. 2018 Jan;21:36–42.

41. Fernandez N, Copenhaver DJ, Vawdrey DK, Kotchoubey H, Stockwell MS. Smartphone Use Among Postpartum Women and Implications for Personal Health Record Utilization. Clin Pediatr . 2017 Apr;56(4):376–81.

42. Filannino M, Stubbs A, Uzuner Ö. Symptom severity prediction from neuropsychiatric clinical records: Overview of 2016 CEGS N-GRID shared tasks Track 2. J Biomed Inform. 2017 Nov;75S:S62–70.

43. First MB, Spitzer RL. The DSM: Not Perfect, but Better Than the Alternative. Psychiatric Times [Internet]. 2003 Apr 1 [cited 2019 Jun 12]; Available from: https://www.psychiatrictimes.com/dsm-not-perfect-better-alternative

44. Fitzpatrick KK, Darcy A, Vierhile M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial [Internet]. Vol. 4, JMIR Mental Health. 2017. p. e19. Available from: http://dx.doi.org/10.2196/mental.7785

45. Fortin J-P, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, et al. Harmonization of multi-site diffusion tensor imaging data. Neuroimage. 2017 Nov 1;161:149–70.

46. Fortney L. Blockchain, Explained [Internet]. Investopedia. 2019 [cited 2019 Jun 5]. Available from: https://www.investopedia.com/terms/b/blockchain.asp

47. Gartlehner G, Hansen RA, Morgan LC, Thaler K, Lux L, Van Noord M, et al. Comparative benefits and harms of second-generation antidepressants for treating major depressive disorder: an updated meta-analysis. Ann Intern Med. 2011 Dec 6;155(11):772–85.

48. Gartner Reprint [Internet]. [cited 2019 Feb 28]. Available from: https://www.gartner.com/doc/reprints?id=1-65WC0O1&ct=190128&st=sb

49. Gelenberg AJ, Freeman MP, Markowitz JC, et al. Practice guideline for the treatment of patients with major depressive disorder third edition. Am J Psychiatry. 2010;167(10):1.

50. Gerra ML, Marchesi C, Amat JA, Blier P, Hellerstein DJ, Stewart JW. Does negative affectivity predict differential response to an SSRI versus a non-SSRI antidepressant? J Clin Psychiatry. 2014 Sep;75(9):e939–44.

51. Gómez D, Rojas A. An Empirical Overview of the No Free Lunch Theorem and Its Effect on Real-World Machine Learning Classification. Neural Comput [Internet]. 2016 Jan;28(1):216–28. Available from: http://dx.doi.org/10.1162/NECO_a_00793

52. Graham J. Artificial Intelligence, Machine Learning, And The FDA. Forbes Magazine [Internet]. 2016 Aug 19 [cited 2019 May 7]; Available from: https://www.forbes.com/sites/theapothecary/2016/08/19/artificial-intelligence-machine-learning-and-the-fda/

53. Greenberg AJ, Haney D, Blake KD, Moser RP, Hesse BW. Differences in Access to and Use of Electronic Personal Health Information Between Rural and Urban Residents in the United States. J Rural Health. 2018 Feb;34 Suppl 1:s30–8.

54. Greiner R, Grove AJ, Roth D. Learning cost-sensitive active classifiers. Artif Intell. 2002;139(2):137–74.

55. Gressin S. The Equifax Data Breach: What to Do [Internet]. Consumer Information. 2017 [cited 2019 May 7]. Available from: https://www.consumer.ftc.gov/blog/2017/09/equifax-data-breach-what-do

56. Grob GN. Origins of DSM-I: a study in appearance and reality. Am J Psychiatry. 1991 Apr;148(4):421–31.

57. Guy W, National Institute of Mental Health (U.S.)., Psychopharmacology Research Branch., Early Clinical Drug Evaluation Program. ECDEU assessment manual for psychopharmacology [Internet]. Rockville, Md.: U.S. Dept. of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, National Institute of Mental Health, Psychopharmacology Research Branch, Division of Extramural Research Programs; 1976. Available from: https://ualberta.worldcat.org/title/ecdeu-assessment-manual-for-psychopharmacology/oclc/2344751

58. Habert J, Katzman MA, Oluboka OJ, McIntyre RS, McIntosh D, MacQueen GM, et al. Functional Recovery in Major Depressive Disorder: Focus on Early Optimized Treatment. Prim Care Companion CNS Disord [Internet]. 2016 Sep 1;18(5). Available from: http://dx.doi.org/10.4088/PCC.15r01926

59. Hajian S, Bonchi F, Castillo C. Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. p. 2125–6. (KDD '16).

60. Halevy A, Norvig P, Pereira F. The Unreasonable Effectiveness of Data [Internet]. Vol. 24, IEEE Intelligent Systems. 2009. p. 8–12. Available from: http://dx.doi.org/10.1109/mis.2009.36

61. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960 Feb;23:56–62.

62. Hassanalieragh M, Page A, Soyata T, Sharma G, Aktas M, Mateos G, et al. Health Monitoring and Management Using Internet-of-Things (IoT) Sensing with Cloud-Based Processing: Opportunities and Challenges. In: 2015 IEEE International Conference on Services Computing. 2015. p. 285–92.

63. Health Canada. Building better access to digital health technologies - Canada.ca [Internet]. 2017 [cited 2019 May 7]. Available from: https://www.canada.ca/en/health-canada/corporate/transparency/regulatory-transparency-and-openness/improving-review-drugs-devices/building-better-access-digital-health-technologies.html

64. Health Canada. Health Portfolio - Canada.ca [Internet]. 2017 [cited 2019 Jun 14]. Available from: https://www.canada.ca/en/health-canada/corporate/health-portfolio.html

65. Health Canada. Notice: Health Canada's Approach to Digital Health Technologies - Canada.ca [Internet]. 2018 [cited 2019 May 7]. Available from: https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/activities/announcements/notice-digital-health-technologies.html

66. Heart T, Ben-Assuli O, Shabtai I. A review of PHR, EMR and EHR integration: A more personalized healthcare and public health policy. Health Policy and Technology. 2017 Mar 1;6(1):20–5.

67. Henkel V, Seemüller F, Obermeier M, Adli M, Bauer M, Mundt C, et al. Does early improvement triggered by antidepressants predict response/remission? Analysis of data from a naturalistic study on a large sample of inpatients with major depression. J Affect Disord. 2009 Jun;115(3):439–49.

68. Henson P, Wisniewski H, Hollis C, Keshavan M, Torous J. Digital mental health apps and the therapeutic alliance: initial review. BJPsych open [Internet]. 2019;5(1). Available from:

https://www.cambridge.org/core/journals/bjpsych-open/article/digital-mental-health-apps-and-the-therapeutic-alliance-initial-review/84D2BF70EEA1EAD7E681FF012651B55E

69. Hippocrates. Works of Hippocrates, Vol. I–IV. (Trans. W. H. S. Jones & E. T. Withington). Cambridge, MA: Harvard University Press; 1923-1931.

70. Horwitz AV, Wakefield JC, Lorenzo-Luaces L. History of Depression. In: The Oxford Handbook of Mood Disorders. 2016. p. 1–24.

71. Houts AC. Fifty years of psychiatric nomenclature: reflections on the 1943 War Department Technical Bulletin, Medical 203. J Clin Psychol. 2000 Jul;56(7):935–67.

72. Iacobucci G. London GP clinic sees big jump in patient registrations after Babylon app launch. BMJ. 2017 Dec 21;359:j5908.

73. Iniesta R, Malki K, Maier W, Rietschel M, Mors O, Hauser J, et al. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. J Psychiatr Res [Internet]. 2016 Jul;78:94–102. Available from: http://dx.doi.org/10.1016/j.jpsychires.2016.03.016

74. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry. 2010 Jul;167(7):748–51.

75. Insel T. Post by Former NIMH Director Thomas Insel: Transforming Diagnosis [Internet]. National Institute of Mental Health. 2013 [cited 2019 Feb 10]. Available from: https://www.nimh.nih.gov/about/directors/thomas-insel/blog/2013/transforming-diagnosis.shtml/index.shtml

76. Iqbal E, Mallah R, Rhodes D, Wu H, Romero A, Chang N, et al. ADEPt, a semantically-enriched pipeline for extracting adverse drug events from free-text electronic health

records [Internet]. Vol. 12, PLOS ONE. 2017. p. e0187121. Available from: http://dx.doi.org/10.1371/journal.pone.0187121

77. Jack. Copeland B. The Essential Turing. Clarendon Press; 2004. 620 p.

78. Jafari M, Safavi-Naini R, Sheppard NP. A rights management approach to protection of privacy in a cloud of electronic health records. In: Proceedings of the 11th annual ACM workshop on Digital rights management. ACM; 2011. p. 23–30.

79. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 May 2;13(6):395–405.

80. Jhamb M, Abdel-Kader K, Yabes J, Wang Y, Weisbord SD, Unruh M, et al. Comparison of fatigue, pain and depression in patients with advanced kidney disease and cancer - symptom burden and clusters. J Pain Symptom Manage [Internet]. 2018 Dec 12; Available from: http://dx.doi.org/10.1016/j.jpainsymman.2018.12.006

81. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017 Dec;2(4):230–43.

82. Kaelber DC, Jha AK, Johnston D, Middleton B, Bates DW. A research agenda for personal health records (PHRs). J Am Med Inform Assoc. 2008 Nov;15(6):729–36.

83. Kalia M, Costa E Silva J. Biomarkers of psychiatric diseases: current status and future prospects. Metabolism. 2015 Mar;64(3 Suppl 1):S11–5.

84. Kanne SM, Carpenter LA, Warren Z. Screening in toddlers and preschoolers at risk for autism spectrum disorder: Evaluating a novel mobile-health screening tool. Autism Res. 2018;11(7):1038–49.

85. Kaplan B. How Should Health Data Be Used?: Privacy, Secondary Use, and Big Data Sales. Camb Q Healthc Ethics. 2016 Apr;25(2):312–29.

86. Kaplan HI. Kaplan & Sadock's Comprehensive Textbook of Psychiatry. Wolters Kluwer Health/Lippincott Williams & Wilkins; 2009. 4520 p.

87. Kennedy SH, Lam RW, McIntyre RS, Tourjman SV, Bhat V, Blier P, et al. Group, CDW, 2016. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 clinical guidelines for the management of adults with major depressive disorder: section 3. Pharmacological treatments. Pharmacological Treatments Can J Psychiatr. 61(9):540–60.

88. Kornhuber J, Terfloth L, Bleich S, Wiltfang J, Rupprecht R. Molecular properties of psychopharmacological drugs determining non-competitive inhibition of 5-HT3A receptors. Eur J Med Chem. 2009 Jun;44(6):2667–72.

89. Kornstein SG, Fava M, Jiang Q, Tourian KA. Analysis of depressive symptoms in patients with major depressive disorder treated with desvenlafaxine or placebo. Psychopharmacol Bull. 2009;42(3):21–35.

90. Kostkova P, Brewer H, de Lusignan S, Fottrell E, Goldacre B, Hart G, et al. Who Owns the Data? Open Data for Healthcare [Internet]. Vol. 4, Frontiers in Public Health. 2016. Available from: http://dx.doi.org/10.3389/fpubh.2016.00007

91. Kostro D, Abdulkadir A, Durr A, Roos R, Leavitt BR, Johnson H, et al. Correction of inter-scanner and within-subject variance in structural MRI based automated diagnosing. Neuroimage. 2014 Sep;98:405–15.

92. Krepel N, Rush AJ, Iseger TA, Sack AT, Arns M. Can psychological features predict antidepressant response to rTMS? A Discovery-Replication approach. Psychol Med. 2019 Jan 24;1–9.

93. Lam RW, Endicott J, Hsu M-A, Fayyad R, Guico-Pabia C, Boucher M. Predictors of functional improvement in employed adults with major depressive disorder treated with desvenlafaxine. Int Clin Psychopharmacol. 2014 Sep;29(5):239–51.

94. Lawlor C. From Melancholia to Prozac: A History of Depression. OUP Oxford; 2012. 265 p.

95. Lee CH, Yoon H-J. Medical big data: promise and challenges. Kidney Res Clin Pract. 2017 Mar;36(1):3–11.

96. Lee Y, Ragguett R-M, Mansur RB, Boutilier JJ, Rosenblat JD, Trevizol A, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. J Affect Disord [Internet]. 2018 Dec 1;241:519–32. Available from: http://dx.doi.org/10.1016/j.jad.2018.08.073

97. Lieblich SM, Castle DJ, Pantelis C, Hopwood M, Young AH, Everall IP. High heterogeneity and low reliability in the diagnosis of major depression will impair the development of new drugs. BJPsych Open. 2015 Oct;1(2):e5–7.

98. Lin HW, Tegmark M, Rolnick D. Why Does Deep and Cheap Learning Work So Well? J Stat Phys. 2017 Sep 1;168(6):1223–47.

99. Lovibond SH, Lovibond PF. Manual for the Depression Anxiety Stress Scales. Psychology Foundation of Australia; 1996. 42 p.

100. Lozupone M, La Montagna M, D'Urso F, Daniele A, Greco A, Seripa D, et al. The Role of Biomarkers in Psychiatry. In: Guest PC, editor. Reviews on Biomarker Studies in Psychiatric and Neurodegenerative Disorders. Cham: Springer International Publishing; 2019. p. 135–62.

101. Malki K, Koritskaya E, Harris F, Bryson K, Herbster M, Tosto MG. Epigenetic differences in monozygotic twins discordant for major depressive disorder. Transl Psychiatry. 2016 Jun 14;6(6):e839.

102. Mallinckrodt CH, Prakash A, Houston JP, Swindle R, Detke MJ, Fava M. Differential antidepressant symptom efficacy: placebo-controlled comparisons of duloxetine and SSRIs (fluoxetine, paroxetine, escitalopram). Neuropsychobiology. 2007 Nov 23;56(2-3):73–85.

103. Marrone S. Understanding barriers to health care: a review of disparities in health care services among indigenous populations. Int J Circumpolar Health. 2007 Jun;66(3):188–98.

104. Masand PS. Tolerability and adherence issues in antidepressant therapy. Clin Ther. 2003 Aug;25(8):2289–304.

105. Masuda K, Nakanishi M, Okamoto K, Kawashima C, Oshita H, Inoue A, et al. Different functioning of prefrontal cortex predicts treatment response after a selective serotonin reuptake inhibitor treatment in patients with major depression. J Affect Disord. 2017 May;214:44–52.

106. McIntyre R, Kennedy S, Bagby RM, Bakish D. Assessing full remission. J Psychiatry Neurosci [Internet]. 2002 Jul;27(4):235–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12174732

107. McLellan RA, Oscarson M, Seidegård J, Evans DA, Ingelman-Sundberg M. Frequent occurrence of CYP2D6 gene duplication in Saudi Arabians. Pharmacogenetics. 1997 Jun;7(3):187–91.

108. Medical Device (SaMD) Working Group. “Software as a Medical Device”: Possible Framework for Risk Categorization and Corresponding Considerations. In International Medical Device Regulators Forum; 2014.

109. Minsky, Marvin (1967), Computation: Finite and Infinite Machines, Englewood Cliffs, N.J.: Prentice-Hall.

110. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry [Internet]. 1979 Apr;134:382–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/444788

111. Moravec HP. The Stanford Cart and the CMU Rover. Proc IEEE. 1983 Jul;71(7):872–84.

112. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013 Apr 3;309(13):1351–2.

113. nbn2r.org - NBN New Knowledge, New Nomenclature [Internet]. [cited 2019 Jun 12]. Available from: http://nbn2r.org/

114. Neuroscience-based Nomenclature [Internet]. European College for Neuropsychopharmacology (ECNP). [cited 2019 Jun 12]. Available from: https://www.ecnp.eu/research-innovation/nomenclature.aspx

115. NIMH » Discussion [Internet]. National Institute of Mental Health. [cited 2019 Jun 12]. Available from: https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/discussion.shtml

116. NIMH » RDoC Matrix [Internet]. National Institute of Mental Health. [cited 2019 Jun 12]. Available from: https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/constructs/rdoc-matrix.shtml

117. Norbury A. Response heterogeneity: Challenges for personalised medicine and big data approaches in psychiatry and chronic pain. F1000Res [Internet]. 2018 [cited 2019 May 7];7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5820606/

118. O'Doherty KC, Christofides E, Yen J, Bentzen HB, Burke W, Hallowell N, et al. If you build it, they will come: unintended future uses of organised health data collections [Internet]. Vol. 17, BMC Medical Ethics. 2016. Available from: http://dx.doi.org/10.1186/s12910-016-0137-x

119. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016 Sep 29;375(13):1216–9.

120. Olfson M, Mojtabai R, Sampson NA, Hwang I, Druss B, Wang PS, et al. Dropout from outpatient mental health care in the United States. Psychiatr Serv. 2009 Jul;60(7):898–907.

121. Olgiati P, Serretti A, Souery D, Dold M, Kasper S, Montgomery S, et al. Early improvement and response to antidepressant medications in adults with major depressive disorder. Meta-analysis and study of a sample with treatment-resistant depression. J Affect Disord. 2018 Feb;227:777–86.

122. Oluboka OJ, Katzman MA, Habert J, McIntosh D, MacQueen GM, Milev RV, et al. Functional Recovery in Major Depressive Disorder: Providing Early Optimal Treatment for the Individual Patient. Int J Neuropsychopharmacol [Internet]. 2018 Feb 1;21(2):128–44. Available from: http://dx.doi.org/10.1093/ijnp/pyx081

123. Patel MJ, Khalaf A, Aizenstein HJ. Studying depression using imaging and machine learning methods. Neuroimage Clin. 2016;10:115-123.

124. Pearl J. Causality by Judea Pearl [Internet]. Cambridge University Press; 2009 [cited 2019 Feb 10]. Available from: https://www.cambridge.org/core/books/causality/B0046844FAE10CBF274D4ACBDAEB5F5B

125. Piatetsky G. Gainers, Losers, and Trends in Gartner 2019 Magic Quadrant for Data Science and Machine Learning Platforms [Internet]. 2019 [cited 2019 Feb 28]. Available from: https://www.kdnuggets.com/2019/02/gartner-2019-mq-data-science-machine-learning-changes.html

126. Pisanu C, Heilbronner U, Squassina A. The Role of Pharmacogenomics in Bipolar Disorder: Moving Towards Precision Medicine [Internet]. Vol. 22, Molecular Diagnosis & Therapy. 2018. p. 409–20. Available from: http://dx.doi.org/10.1007/s40291-018-0335-y

127. Quiñonero-Candela J. Dataset Shift in Machine Learning. MIT Press; 2009. 229 p.

128. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine. 2018 May 8;1(1):18.

129. RapidMiner. Lightning Fast Data Science Platform for Teams | RapidMiner© [Internet]. RapidMiner. RapidMiner; 2016 [cited 2019 Mar 1]. Available from: https://rapidminer.com/

130. Razzaki S, Baker A, Perov Y, Middleton K, Baxter J, Mullarkey D, et al. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis [Internet]. arXiv [cs.AI]. 2018. Available from: http://arxiv.org/abs/1806.10698

131. Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, et al. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. Am J Psychiatry [Internet]. 2013 Jan;170(1):59–70. Available from: http://dx.doi.org/10.1176/appi.ajp.2012.12070999

132. Rizvi SL, Dimeff LA, Skutch J, Carroll D, Linehan MM. A pilot study of the DBT coach: an interactive mobile phone application for individuals with borderline personality disorder and substance use disorder. Behav Ther. 2011 Dec;42(4):589–600.

133. Roehrs A, da Costa CA, da Rosa Righi R. OmniPHR: A distributed architecture model to integrate personal health records. J Biomed Inform. 2017 Jul;71:70–81.

134. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958 Nov;65(6):386–408.

135. Ross JS, Lehman R, Gross CP. The importance of clinical trial data sharing: toward more open science. Circ Cardiovasc Qual Outcomes. 2012 Mar 1;5(2):238–40.

136. Rumelhart DE, Hinton GE, Williams RJ, et al. Learning representations by back-propagating errors. Cognitive modeling. 1988;5(3):1.

137. Rush AJ, Kraemer HC, Sackeim HA, Fava M, Trivedi MH, Frank E, Ninan PT, Thase ME, Gelenberg AJ, Kupfer DJ, Regier DA, Rosenbaum JF, Ray O, Schatzberg AF. Report by the ACNP Task Force on response and remission in major depressive disorder. Neuropsychopharmacology. 2006;31(9):1841–1853.

138. Sano A, Phillips AJ, Yu AZ, McHill AW, Taylor S, Jaques N, et al. Recognizing Academic Performance, Sleep Quality, Stress Level, and Mental Health using Personality Traits, Wearable Sensors and Mobile Phones. Int Conf Wearable Implant Body Sens Netw [Internet]. 2015 Jun;2015. Available from: http://dx.doi.org/10.1109/BSN.2015.7299420

139. SAS. Machine Learning: What it is and why it matters [Internet]. SAS Analytics Insights. 2019 [cited 2019 Jun 13]. Available from: https://www.sas.com/en_ca/insights/analytics/machine-learning.html

140. Sejnowski TJ, Rosenberg CR. Parallel networks that learn to pronounce English text. Complex Systems. 1987;1(1):145–68.

141. Settles B. Active learning literature survey [Internet]. University of Wisconsin-Madison Department of Computer Sciences; 2009. Available from: https://minds.wisconsin.edu/handle/1793/60660

142. Sevel LS, Boissoneault J, Letzen JE, Robinson ME, Staud R. Structural brain changes versus self-report: machine-learning classification of chronic fatigue syndrome patients. Exp Brain Res. 2018 Aug;236(8):2245–53.

143. Shamir D, Szor H, Melamed Y. Dropout, early termination and detachment from a public psychiatric clinic. Psychiatr Danub. 2010 Mar;22(1):46–50.

144. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med. February 2019:1-23.

145. Sheehan DV, Nakagome K, Asami Y, Pappadopulos EA, Boucher M. Restoring function in major depressive disorder: A systematic review. J Affect Disord [Internet]. 2017 Jun;215:299–313. Available from: http://dx.doi.org/10.1016/j.jad.2017.02.029

146. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform. 2018 Sep;22(5):1589–604.

147. Shorter E. The history of nosology and the rise of the Diagnostic and Statistical Manual of Mental Disorders. Dialogues Clin Neurosci. 2015 Mar;17(1):59–67.

148. Shortliffe EH, Buchanan BG. A Model of Inexact Reasoning in Medicine. Math Biosci. 1975 Apr 1;23(3-4):351–79.

149. Simon GE, Johnson E, Lawrence JM, Rossom RC, Ahmedani B, Lynch FL, et al. Predicting Suicide Attempts and Suicide Deaths Following Outpatient Visits Using Electronic Health Records. Am J Psychiatry. 2018 Oct 1;175(10):951–60.

150. Sittig DF, Singh H. Legal, ethical, and financial dilemmas in electronic health record adoption and use. Pediatrics. 2011 Apr;127(4):e1042–7.

151. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, et al. Grand challenges in clinical decision support. J Biomed Inform. 2008 Apr;41(2):387–92.

152. Smith K. Mental health: a world of depression. Nature. 2014 Nov 13;515(7526):181.

153. Smith K. Mental health: a world of depression. Nature. 2014 Nov 13;515(7526):181.

154. Soares CN, Endicott J, Boucher M, Fayyad RS, Guico-Pabia CJ. Predictors of functional response and remission with desvenlafaxine 50 mg/d in patients with major depressive disorder. CNS Spectr. 2014 Dec;19(6):519–27.

155. Soni D. Dealing with Imbalanced Classes in Machine Learning [Internet]. Towards Data Science. Towards Data Science; 2018 [cited 2019 May 7]. Available from: https://towardsdatascience.com/dealing-with-imbalanced-classes-in-machine-learning-d43d6fa19d2

156. Spencer K, Sanders C, Whitley EA, Lund D, Kaye J, Dixon WG. Patient Perspectives on Sharing Anonymized Personal Health Data Using a Digital System for Dynamic Consent and Research Feedback: A Qualitative Study. J Med Internet Res. 2016 Apr 15;18(4):e66.

157. Straube S, Krell MM. How to evaluate an agent's behavior to infrequent events?-Reliable performance estimation insensitive to class distribution. Front Comput Neurosci. 2014 Apr 10;8:43.

158. Sung SC, Wisniewski SR, Luther JF, Trivedi MH, Rush AJ, COMED Study Team. Pre-treatment insomnia as a predictor of single and combination antidepressant outcomes: a CO-MED report. J Affect Disord. 2015 Mar 15;174:157–64.

159. Terry AL, Stewart M, Fortin M, Wong ST, Kennedy M, Burge F, et al. Gaps in primary healthcare electronic medical record research and knowledge: findings of a pan-Canadian study. Healthc Policy. 2014;10(1):46–59.

160. Torous J, Larsen ME, Depp C, Cosco TD, Barnett I, Nock MK, et al. Smartphones, Sensors, and Machine Learning to Advance Real-Time Prediction and Interventions for Suicide Prevention: a Review of Current Progress and Next Steps. Curr Psychiatry Rep. 2018 Jun 28;20(7):51.

161. Torous JB, Chan SR, Gipson SY-MT, Kim JW, Nguyen T-Q, Luo J, et al. A Hierarchical Framework for Evaluation and Informed Decision Making Regarding Smartphone Apps for Clinical Care. Psychiatr Serv. 2018 May 1;69(5):498–500.

162. Torres SJ, Nowson CA, Worsley A. Dietary electrolytes are related to mood. Br J Nutr [Internet]. 2008 Nov;100(5):1038–45. Available from: http://dx.doi.org/10.1017/S0007114508959201

163. Tran T, Kavuluru R. Predicting mental conditions based on “history of present illness” in psychiatric notes with deep neural networks. J Biomed Inform. 2017 Nov 1;75:S138–48.

164. Trivedi MH, Morris DW, Grannemann BD, Mahadi S. Symptom clusters as predictors of late response to antidepressant treatment. J Clin Psychiatry [Internet]. 2005 Aug;66(8):1064–70. Available from: https://www.ncbi.nlm.nih.gov/pubmed/16086624

165. Trivedi MH, Rush AJ, Wisniewski SR, Nierenberg AA, Warden D, Ritz L, et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. Am J Psychiatry [Internet]. 2006 Jan;163(1):28–40. Available from: http://dx.doi.org/10.1176/appi.ajp.163.1.28

166. U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER). Major Depressive Disorder: Developing Drugs for Treatment, Guidance for Industry [Internet]. 2018. Available from: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM611259.pdf
167. U.S. Food and Drug Administration. About FDA [Internet]. U.S. Food and Drug Administration (FDA). 2019b [cited 2019 Jun 14]. Available from: http://www.fda.gov/about-fda
168. U.S. Food and Drug Administration. Digital Health Innovation Action Plan. 2018.
169. U.S. Food and Drug Administration. Digital health software precertification (Pre-Cert) program. Available at:)(Accessed May 27, 2018) https://www fda gov/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/default htm View in Article. 2018b.
170. US FDA Artificial Intelligence and Machine Learning Discussion Paper; 2019. Available from: https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/SoftwareasaMedicalDevice/UCM635052.pdf
171. US FDA Artificial Intelligence and Machine Learning Discussion Paper. Available from: https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/SoftwareasaMedicalDevice/UCM635052.pdf
172. van de Meent J-W. Unsupervised Machine Learning and Data Mining [Internet]. DS 5230 / DS 4420 Class Notes; 2018; Northeastern University. Available from: https://course.ccs.neu.edu/ds5230f18/assets/slides/ds5230-f18-lecture-01.pdf

173. Vassos E, Agerbo E, Mors O, Pedersen CB. Urban–rural differences in incidence rates of psychiatric disorders in Denmark. Br J Psychiatry. 2016 May;208(5):435–40.

174. Wen J, Hassanpour N, Greiner R. Weighted Gaussian Process for Estimating Treatment Effect. 30th Conference on Neural Information Processing Systems [Internet]. 2016; Available from: https://pdfs.semanticscholar.org/9706/6ad9f744213f54e1afa04c991055b1345f03.pdf

175. Wikipedia contributors. AI winter [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Jul 4]. Available from: https://en.wikipedia.org/w/index.php?title=AI_winter&oldid=903461097

176. Wikipedia contributors. Arthur Samuel [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Jul 4]. Available from: https://en.wikipedia.org/w/index.php?title=Arthur_Samuel&oldid=887136586

177. Wikipedia contributors. Blockchain [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Jun 5]. Available from: https://en.wikipedia.org/w/index.php?title=Blockchain&oldid=900245799

178. Wikipedia contributors. Confusion matrix [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Apr 24]. Available from: https://en.wikipedia.org/w/index.php?title=Confusion_matrix&oldid=881721342

179. Wikipedia contributors. Diagnostic and Statistical Manual of Mental Disorders [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Jun 12]. Available from: https://en.wikipedia.org/w/index.php?title=Diagnostic_and_Statistical_Manual_of_Mental_Disorders&oldid=898942134

180. Wikipedia contributors. Grace Hopper [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Jul 4]. Available from: https://en.wikipedia.org/w/index.php?title=Grace_Hopper&oldid=904577918

181. Wikipedia contributors. Humorism [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Jun 7]. Available from: https://en.wikipedia.org/w/index.php?title=Humorism&oldid=899951997

182. Wikipedia contributors. Internet of things [Internet]. Wikipedia, The Free Encyclopedia. 2019 [cited 2019 Jun 5]. Available from: https://en.wikipedia.org/w/index.php?title=Internet_of_things&oldid=900133956

183. World Health Organization (WHO). ICD-10 Version:2016 [Internet]. 2016 [cited 2019 Jun 12]. Available from: https://icd.who.int/browse10/2016/en

184. World Health Organization. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. Geneva: World Health Organization; 1992.

185. Zapata BC, Niñirola AH, Fernández-Alemán JL, Toval A. Assessing the privacy policies in mobile personal health records. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2014. p. 4956–9.

186. Zhu X, Vondrick C, Fowlkes C, Ramanan D. Do We Need More Training Data? [Internet]. arXiv [cs.CV]. 2015. Available from: http://arxiv.org/abs/1503.01508

Appendices

Appendix 1. Code for Chapter 2 Concure Classifier

A1.1 Data Preparation

```
#!/usr/bin/env python3

import pandas as pd
import os, scipy.stats
import numpy as np
import pathlib

def Misc():

    #VARIABLES

    #Encode a categorical variable as an integers (not onehot or dummy)
    demow['ETHNIC']=pd.Categorical(demow['ETHNIC']).codes

    #Quick way to encode/binarize
    df.set_index('PATIENT')
    df['ones']=1
    df.pivot(columns='MEDGNX', values='ones')
    df=df.fillna(value=0)

    #Recode a variable
    df['Severity of illness']=df['Severity of illness'].replace(to_replace='Borderline ill',
value='Borderline mentally ill')

    #PIVOTS & JOINS

    #pivots table
    d60p= d60.pivot(index='PATIENT',columns='TESTS',values='VALN')

    #pivot each column of 'lab' to its own table (lab1, lab2, etc)
    lab1= lab.pivot(index='PATIENT',columns='LPARM',values='LVALN')

    #relabels columns so join works
    lab1.columns='lab1-'+lab1.columns

    #joins tables
    labs=lab1.join([lab2, lab3])

    #Returns vals in df1 that are ALSO in df2
```

```
    np.intersect1d(df1['PATIENT'],df2['PATIENT'])

    #Returns vals in df1 that are NOT in df2
    np.setdiff1d(df1['PATIENT'],df2['PATIENT'])

    #Remove a study's participants (e.g. if missing data)
    dfs3=dfs2.drop(labels=df.index, errors='ignore')


    #MISC

    #Descriptive statistics: central tendency, dispersion and shape
    df['HAMD Total'].describe()

    #drops duplicates if both column 1 AND column 2 have the same row value
    lab=lab.drop_duplicates(subset=['PATIENT', 'LPARM'], keep='first')

    #Quick way to sample a class to even out classes
    a=labels.columns[0]
    b=labels[a].value_counts()
    df2=labels.loc[labels[a]==0]
    df3=labels.loc[labels[a]==1]
    df4=df2.sample(min(b))
    df5=df3.sample(min(b))
    df6=[df4, df5]
    df7=pd.concat(df6)
    df8=df7.sort_index()

    #Quick way to cut data to labels
    data=data[data.index.isin(labels.index)].sort_index()


    #Sample for holdout set at same frequency of labels as dataset
    #Using ~ to cut the holdout set out of the main set, or not using it to get just the holdout set on
its own


holdout=labels.groupby('HAM-D17 1=REMIT').apply(pd.DataFrame.sample,
frac=0.1).reset_index(level='HAM-D17 1=REMIT', drop=True).sort_index()

holdout_labels=labels[data.index.isin(holdout.index)].sort_index()
labels_excluding_holdout=labels[~data.index.isin(holdout.index)].sort_index()
holdout_data=data[data.index.isin(holdout.index)].sort_index()
data_excluding_holdout=data[~data.index.isin(holdout.index)].sort_index()
```

```
holdout_labels.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/refined-combined-study/holdout-labels.csv', index_label='PATIENT')
labels_excluding_holdout.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/refined-combined-study/labels_excluding_holdout.csv', index_label='PATIENT')
holdout_data.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/refined-combined-study/holdout-data.csv', index_label='PATIENT')
data_excluding_holdout.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/refined-combined-study/data_excluding_holdout.csv', index_label='PATIENT')

    #Cgi ordered categories, -1 indicates NaN
    df=df.set_index('PATIENT')
    cat=['Normal, not at all ill',
    'Borderline mentally ill',
    'Mildly ill',
    'Moderately ill',
    'Markedly ill',
    'Severely ill',
    'Among the most extremely ill']
    df['CGI SEVERITY']=pd.Categorical(df['Severity of illness'], categories=cat).codes

    df[df['CGI SEVERITY']==-1]
    >>>output: 89CHQS
    df=df.drop(['89CHQS'])

    #CGI process
    info=pd.read_csv('/media/james/ext4data1/current/projects/pfizer/combined-study/3151a1-3364-csv/deid_cgi.csv', encoding='utf-8')

    info=info[info['CPENM']=='BASELINE DAY -1']

    info=info.drop_duplicates(subset='PATIENT', keep='first')

    data= info.pivot(index='PATIENT',columns='TESTS',values='VALX')

    data.to_csv(path_or_buf='/media/james/ext4data1/current/projects/pfizer/combined-study/3151a1-3364-csv/deid_cgi_ready.csv',index_label='PATIENT')

    #demow ordered sex/race
    cat=['Male', 'Female']

    #split off a variable
    dftherdur=df['THERDUR'].to_frame()
    del df['THERDUR']

    #ethnicity ordering- alphabetical like NIH
    eth=['American Indian or Alaska Native',
```

```
    'Asian',
    'Black or African American',
    'Hispanic or Latino',
    'Middle Eastern or North African',
    'Native Hawaiian or Other Pacific Islander',
    'Other',
    'White']

   #sex ordering- alphabetical
   sex=['Female','Male']

   #ethnic recode
   df1['ETHNIC RECODE']=df1['ETHNIC RECODE'].replace({
'American Indian or Alaska Native':'American Indian or Alaska Native',
'Arabic':'Middle Eastern or North African',
'Asian':'Asian',
'Black':'Black or African American',
'Black or African American':'Black or African American',
'Chinese':'Asian',
'Hispanic':'Hispanic or Latino',
'Indian':'Asian',
'Korean':'Asian',
'Native American':'American Indian or Alaska Native',
'Native Hawaiian or Other Pacific Islander':'Native Hawaiian or Other Pacific Islander',
'Oriental(Asian)':'Asian',
'Other':'Other',
'Other: (Mixed race)':'Other',
'Other: Alaskan Native':'American Indian or Alaska Native',
'Other: Coloured.':'Black or African American',
'Other: Mid eastern':'Middle Eastern or North African',
'Other: Middle eastern':'Middle Eastern or North African',
'Other: Mixed':'Other',
'Other: Mixed race.':'Other',
'Other: Mixed.':'Other',
'Other: Panaminian.':'Hispanic or Latino',
'Other: Russian':'White',
'Other: XXXXXXXXXX':'Other',
'Other:Bi-Racial.':'Other',
'Other:Brazilian':'Hispanic or Latino',
'Other:Cauc/Asian-pacific islander':'Other',
'Other:INDIAN':'Asian',
'Other:Mixed':'Other',
'Other:Mixed race.':'Other',
'Other:Mixed.':'Other',
'Other:Slavic':'White',
'Taiwanese':'Asian',
```

```
 'White':'White'})

    return

def Labeler():
    hamd=pd.read_csv('/media/james/ext4data1/current/projects/pfizer/303-
data/deid_hamd17a.csv')
    df['HAMD 1=REMIT']=np.where(df['HAMD-17 questions Total score derived']<=7, 1, 0)

    df.to_csv(path_or_buf='/media/james/ext4data1/current/projects/pfizer/combined-study/class-
labels.csv', index_label='PATIENT')

    return


def GroupDefiner():
    labels=pd.read_csv('/media/james/ext4data1/current/projects/pfizer/labels-d60-placebo-
remitters.csv', encoding='utf-8').set_index('PATIENT').sort_index()
    placebos=pd.read_csv('/media/james/ext4data1/current/projects/pfizer/placebo-
patients.csv').set_index('PATIENT').sort_index()
    therapy= pd.read_csv('/media/james/ext4data1/current/projects/pfizer/therapy-60-
completed.csv').set_index('PATIENT').sort_index()

    placebos=placebos[placebos['TPNAME']=='Placebo']

    therapy=therapy[therapy['THERDUR>=60']==1]

    final= labels.join([placebos, therapy], how='inner')

    del final['TPNAME']
    del final['THERDUR>=60']

    final.to_csv(path_or_buf='/media/james/ext4data1/current/projects/pfizer/labels-final.csv',
index_label='PATIENT')

    return


def Homeopathy():
    #Cuts all tables to subjects in labels-final
    patients=pd.read_csv('/media/james/ext4data1/current/projects/pfizer/labels-final.csv',
encoding='utf-8').set_index('PATIENT').index

    path= '/media/james/ext4data1/current/projects/pfizer/303-data-baseline/'
    csvs= os.listdir(path)
    for i in csvs:
```

```
        a= pd.read_csv(path+i)
        b= a[a['PATIENT'].isin(patients)]
        b= b.set_index('PATIENT')
        b.to_csv(path_or_buf='/media/james/ext4data1/current/projects/pfizer/303-data-
baseline/cut-'+str(i), index_label='PATIENT')

    return

#>>>
#NOW /CUT/ENCODE THE TABLES DOWN MANUALLY
#>>>

def Binarizer():
    #Use if you're making binarized variables
    csv= ['deid_adverse', 'deid_aemeddra', 'deid_medhist', 'deid_medhist2', 'deid_nsmed',
'deid_othtrt']
    for i in csv:
        info=pd.read_csv('/media/james/ext4data1/current/projects/pfizer/3151A1-303-
csv/'+str(i)+'.csv', encoding='utf-8')
        a= info.set_index(['PATIENT'])
        b= pd.get_dummies(a)
        d= {}
        for j in list(set(b.index)):
            d[j]= b.loc[j].values.flatten()

        maxlen=len(d[max(d, key=lambda k: len(d[k]))])
        for m in d:
            d[m]=np.append(d[m], [0]*(maxlen-len(d[m])))
        d= pd.DataFrame.from_dict(d, orient='index')
        d.columns=list(b.columns)*scipy.stats.mode(b.index).count[0]

d.to_csv(path_or_buf='/media/james/ext4data1/current/projects/pfizer/vecs/vecs_'+str(i)+'.csv',
index_label='PATIENT')
        #this gives a dataframe with all variables binarized

    return


def Harvester():
    '"Because it's a combine. Aha. Ha."'
    #But seriously, joins all tables together by patient row

    info= pd.read_csv('/media/james/ext4data1/current/projects/pfizer/labels-d60-placebo-
remitters.csv', encoding='utf-8').set_index('PATIENT').drop('GROUPLABEL', axis=1)

    path= '/media/james/ext4data1/current/projects/pfizer/303-data-baseline-final/'
```

```
    csvs= os.listdir(path)
    for i in csvs:
        a=pd.read_csv(path+i).set_index('PATIENT')
        info=info.join(a, how='inner')

    info.to_csv(path_or_buf='/media/james/ext4data1/current/projects/pfizer/joined-vecs.csv',
index_label='PATIENT')

    return

def CombineStudies():
    '''Combines all studies in a directory with the same column headers'''

    basedir=input('Click and drag DIRECTORY here: ')
    root=basedir.strip('\' ')
    dirname= os.path.basename(root)

    basefiles=[]

    for path, subdirs, files in os.walk(root):
        for name in files:
            fpath= os.path.join(path, name)
            basefiles=basefiles+[fpath]

    combinedframe= pd.DataFrame()

    for i in basefiles:
        print(i)
        data=pd.read_csv(i, encoding='utf-8').set_index('PATIENT')
        combinedframe=pd.concat([combinedframe,data])

    combinedframe.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/refined-
combined-study/Data/'+dirname+'.csv')
```

A1.2 Missing Value Imputation

```
#!/usr/bin/env python3

import pandas as pd
import os, scipy.stats
import numpy as np
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import MinMaxScaler

def Impute():
```

```
    a=input('Click and drag DATASET WITH MISSING VALUES file here: ')
    a=a.strip('\' ')
    data=pd.read_csv(a, encoding='utf-8').set_index('PATIENT')

    b=input('Click and drag LABELS file here: ')
    b=b.strip('\' ')
    labels=pd.read_csv(b, encoding='utf-8').set_index('PATIENT')

    data= data.dropna(axis='columns', how='all')



    X= SimpleImputer().fit_transform(data)

    mms= MinMaxScaler()
    X2= mms.fit_transform(X)
    X3=pd.DataFrame(data=X2, columns=data.columns, index=data.index)

    X3.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/refined-combined-
study/data.csv', index_label='PATIENT')

    return
```

A1.3 Cross-Validation

```
#!/usr/bin/env python3

import pickle
import numpy as np
import pandas as pd
from sklearn.model_selection import StratifiedKFold

def OuterCv():
    a=input('Click and drag ENTIRE DATASET file here: ')
    a=a.strip('\' ')
    data=pd.read_csv(a, encoding='utf-8').set_index('PATIENT')

    b=input('Click and drag LABELS file here: ')
    b=b.strip('\' ')
    labels=pd.read_csv(b, encoding='utf-8').set_index('PATIENT')

    X= data
    y= np.array(labels[labels.columns[0]])

    train, test= [],[]
```

```
    outer_cv= {'train': [],
               'test': []}

    skf = StratifiedKFold(n_splits=5)
    for train_index, test_index in skf.split(X,y):
        train= X.index[train_index]
        test= X.index[test_index]
        outer_cv['train'].append(train)
        outer_cv['test'].append(test)

    with open('/media/james/ext4data/current/projects/pfizer/combined-study/outer_cv.pickle',
'wb') as f: pickle.dump(outer_cv, f, pickle.HIGHEST_PROTOCOL)

    return

def InnerCv():
    a=input('Click and drag ENTIRE DATASET file here: ')
    a=a.strip('\' ')
    data=pd.read_csv(a, encoding='utf-8').set_index('PATIENT')

    b=input('Click and drag LABELS file here: ')
    b=b.strip('\' ')
    labels=pd.read_csv(b, encoding='utf-8').set_index('PATIENT')

    a=input('Click and drag OUTER CV file here: ')
    a=a.strip('\' ')
    with open(a, 'rb') as f: outer_cv= pickle.load(f)

    foldcount= 0
    for i in range(len(outer_cv['train'])):
        foldcount= foldcount+1

        subjects=pd.DataFrame(index=outer_cv['train'][i])
        X= subjects.join(data)
        y= subjects.join(labels)

        train, test, holdout= [],[],[]
        inner_cv= {'train': [],
                   'test': [],
                   'holdout':[]}

        skf = StratifiedKFold(n_splits=5)
        for train_index, test_index in skf.split(X,y):
            train= X.index[train_index]
            test= X.index[test_index]
```

```
        inner_cv['train'].append(train)
        inner_cv['test'].append(test)

    with open('/media/james/ext4data/current/projects/pfizer/combined-
study/inner_cv_fold_'+str(foldcount)+'.pickle', 'wb') as f: pickle.dump(inner_cv, f,
pickle.HIGHEST_PROTOCOL)

  return
```

A1.4 Feature Selection

```
#!/usr/bin/env python3

from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import SelectFromModel
from sklearn.feature_selection import RFECV
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LassoLarsIC, LassoCV, ElasticNet

import copy, pickle
import numpy as np
import pandas as pd
import itertools

#To flatten feature list and get frequencies:
#[item for sublist in inner_cv['Feature Indices'] for item in sublist]
#from collections import Counter
#b=dict(Counter(a))

def OuterFeats():
  a=input('Click and drag ENTIRE DATASET file here: ')
  a=a.strip('\' ')
  data=pd.read_csv(a, encoding='utf-8').set_index('PATIENT')

  b=input('Click and drag LABELS file here: ')
  b=b.strip('\' ')
  labels=pd.read_csv(b, encoding='utf-8').set_index('PATIENT')

  c=input('Click and drag OUTER CV file here: ')
  c=c.strip('\' ')
  with open(c, 'rb') as f: outer_cv= pickle.load(f)

  folds= len(outer_cv['train'])
  feats=[[0]]*folds
```

```
    for i in range(folds):
        subjects=pd.DataFrame(index=outer_cv['train'][i])
        X= subjects.join(data)
        y= subjects.join(labels)

        llic= SelectFromModel(LassoLarsIC())
        #llic= SelectFromModel(LassoLarsIC(criterion='bic'))

        llic.fit(X,y)
        feats[i]=llic.get_support(indices=True)

    featlist= list(set.intersection(*map(set,feats)))
    featlist.sort()
    feature_csv= pd.DataFrame(index=featlist, data= list(data.columns[featlist]))
    feature_csv.index.name='Feature #'
    feature_csv.columns=['Feature Name']

    print(len(featlist))

    data_cut= data[data.columns[featlist]]

    data_cut.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/combined-
study/data-cut-to-feature-set.csv', index_label='PATIENT')
    feature_csv.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/combined-
study/intersecting-features-index.csv')

    return

def InnerFeats():
    a=input('Click and drag ENTIRE DATASET file here: ')
    a=a.strip('\' ')
    data=pd.read_csv(a, encoding='utf-8').set_index('PATIENT')

    b=input('Click and drag LABELS file here: ')
    b=b.strip('\' ')
    labels=pd.read_csv(b, encoding='utf-8').set_index('PATIENT')

    c=input('Click and drag SINGLE FOLD INNER CV file here: ')
    c=c.strip('\' ')
    with open(c, 'rb') as f: inner_cv= pickle.load(f)

    folds= len(inner_cv['train'])
    thisfold=input('Which # fold is this? ')
    feats=[[0]]*folds

    #This is correct because we are mimicking the entire L(.) procedure as if D-1 were D.
```

```
    for i in range(folds):
        subjects=pd.DataFrame(index=inner_cv['train'][i])
        X= subjects.join(data)
        y= subjects.join(labels)

        #llic= SelectFromModel(LassoLarsIC(criterion='bic'))
        llic= SelectFromModel(LassoLarsIC())
        llic.fit(X,y)
        feats[i]=llic.get_support(indices=True)

    featlist= list(set.intersection(*map(set,feats)))
    featlist.sort()
    feature_csv= pd.DataFrame(index=featlist, data= list(data.columns[featlist]))
    feature_csv.index.name='Feature #'
    feature_csv.columns=['Feature Name']

    print(len(featlist))

    data_cut= data[data.columns[featlist]]

    data_cut.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/combined-
study/data-cut-to-feature-set-for-inner-fold-'+str(thisfold)+'.csv', index_label='PATIENT')
    feature_csv.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/combined-
study/intersecting-features-index-for-inner-fold-'+str(thisfold)+'.csv')

    return

def HoldoutCut():
    a=input('Click and drag FEATURE SELECTED ENTIRE DATASET file here: ')
    a=a.strip('\' ')
    data=pd.read_csv(a, encoding='utf-8').set_index('PATIENT')

    b=input('Click and drag HOLDOUT DATA file here: ')
    b=b.strip('\' ')
    hdata=pd.read_csv(b, encoding='utf-8').set_index('PATIENT')

    data_cut=hdata[data.columns]

    data_cut.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/combined-
study/holdout-data-cut-to-feature-set.csv', index_label='PATIENT')

'''
In case we need to re-integrate individual folds and run through them five at a time:

    foldgroup=[]
    for i in range(0, len(feats), 5):
```

```
        foldgroup.append(feats[i:i+5])

    for i in range(len(foldgroup)):
        featlist= list(set.intersection(*map(set,foldgroup[i])))
        featlist.sort()

        data_cut= data[data.columns[featlist]]
'''
```

A1.5 Model Training, Selection, and Testing

```
#!/usr/bin/env python3

import numpy as np
import pandas as pd
import copy, pickle
from sklearn import svm, naive_bayes, neighbors, ensemble, linear_model, tree, neural_network

def EntireDataset():
    a=input('Click and drag FEATURE SELECTED ENTIRE DATASET file here: ')
    a=a.strip('\' ')
    data=pd.read_csv(a, encoding='utf-8').set_index('PATIENT')

    b=input('Click and drag LABELS file here: ')
    b=b.strip('\' ')
    labels_df=pd.read_csv(b, encoding='utf-8').set_index('PATIENT')
    labels=np.array(labels_df[labels_df.columns[0]])

    nfeatsmax= len(data.columns)
    nfeatsneural= round((nfeatsmax*2/3))

    rf= ensemble.RandomForestClassifier(max_features=nfeatsmax,
max_depth=5,bootstrap=False)
    et= ensemble.ExtraTreesClassifier(max_features=nfeatsmax, max_depth=5, bootstrap=False)
    kn= neighbors.KNeighborsClassifier(n_neighbors=nfeatsmax, p=1)
    nb= naive_bayes.GaussianNB()
    dt= tree.DecisionTreeClassifier(max_features=nfeatsmax, max_depth=5, criterion='entropy')
    ls= svm.LinearSVC(penalty='l1', dual=False)
    gb= ensemble.GradientBoostingClassifier(loss='exponential', max_depth=2)
    nn=
neural_network.MLPClassifier(hidden_layer_sizes=(nfeatsneural,nfeatsneural,nfeatsneural,),
learning_rate_init=0.0001, max_iter=500)
    ab= ensemble.AdaBoostClassifier()
    bc= ensemble.BaggingClassifier(base_estimator=rf)
    vc= ensemble.VotingClassifier(estimators=[('gb', gb),('ab', ab),('bc', bc)], voting='soft')
```

```
    estimators= {#'randomforest': rf,
                 #'extratrees': et,
                 #'kneighbors': kn,
                 #'naivebayes': nb,
                 #'decisiontree': dt,
                 'linearsvc': ls,
                 #'gboost': gb,
                 #'neuralnet': nn,
                 #'adaboost': ab,
                 #'bagging': bc,
                 #'voting': vc,
                 }

    results= {'estimator':[],
              'subjects':[],
              'labels':[],
              'predictions':[],
              'scores':[],
              'attempts':[]}

    for j,k in zip(estimators.keys(), estimators.values()):
        k.fit(data, labels)
        predict_train= k.predict(data)
        train_scores= [1 if x==y else 0 for x,y in zip(labels, predict_train)]
        results['estimator'].extend([j]*len(data))
        results['subjects'].extend(data.index)
        results['labels'].extend(labels)
        results['predictions'].extend(predict_train)
        results['scores'].extend(train_scores)
        results['attempts'].extend([1]*len(data))

    results_df=pd.DataFrame.from_dict(results).set_index('subjects')
    results_df.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/combined-
study/entire_dataset_results.csv')

    with open('/media/james/ext4data/current/projects/pfizer/combined-
study/trainedclassifier.pickle', 'wb') as f: pickle.dump(k, f, pickle.HIGHEST_PROTOCOL)

    print('ENTIRE DATASET ACCURACY')
    trd= results_df.groupby('estimator').sum()
    trsum= (trd['scores']/trd['attempts'])*100
    print(trsum)

    return
```

```
def HoldoutDataset():
    a=input('Click and drag HOLDOUT FEATURE SELECTED DATA file here: ')
    a=a.strip('\' ')
    data=pd.read_csv(a, encoding='utf-8').set_index('PATIENT')

    b=input('Click and drag HOLDOUT LABELS file here: ')
    b=b.strip('\' ')
    labels_df=pd.read_csv(b, encoding='utf-8').set_index('PATIENT')
    labels=np.array(labels_df[labels_df.columns[0]])

    a=input('Click and drag TRAINED CLASSIFIER file here: ')
    a=a.strip('\' ')
    with open(a, 'rb') as f: k= pickle.load(f)

    results= {'estimator':[],
              'subjects':[],
              'labels':[],
              'predictions':[],
              'scores':[],
              'attempts':[]}

    j=input('Type NAME of classifier: ')
    predict_train= k.predict(data)
    train_scores= [1 if x==y else 0 for x,y in zip(labels, predict_train)]
    results['estimator'].extend([j]*len(data))
    results['subjects'].extend(data.index)
    results['labels'].extend(labels)
    results['predictions'].extend(predict_train)
    results['scores'].extend(train_scores)
    results['attempts'].extend([1]*len(data))

    results_df=pd.DataFrame.from_dict(results).set_index('subjects')
    results_df.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/combined-
study/holdout_dataset_results.csv')

    print('HOLDOUT DATASET ACCURACY')
    trd= results_df.groupby('estimator').sum()
    trsum= (trd['scores']/trd['attempts'])*100
    print(trsum)

    return

def OuterFolds():
    a=input('Click and drag FEATURE SELECTED ENTIRE DATASET file here: ')
    a=a.strip('\' ')
    data=pd.read_csv(a, encoding='utf-8').set_index('PATIENT')
```

```
    b=input('Click and drag LABELS file here: ')
    b=b.strip('\' ')
    labels=pd.read_csv(b, encoding='utf-8').set_index('PATIENT')

    c=input('Click and drag OUTER CV file here: ')
    c=c.strip('\' ')
    with open(c, 'rb') as f: outer_cv= pickle.load(f)

    folds= len(outer_cv['train'])
    nfeatsmax= len(data.columns)
    nfeatsneural= round((nfeatsmax*2/3))

    rf= ensemble.RandomForestClassifier(max_features=nfeatsmax,
max_depth=5,bootstrap=False)
    et= ensemble.ExtraTreesClassifier(max_features=nfeatsmax, max_depth=5, bootstrap=False)
    kn= neighbors.KNeighborsClassifier(n_neighbors=nfeatsmax, p=1)
    nb= naive_bayes.GaussianNB()
    dt= tree.DecisionTreeClassifier(max_features=nfeatsmax, max_depth=5, criterion='entropy')
    ls= svm.LinearSVC(penalty='l1', dual=False)
    gb= ensemble.GradientBoostingClassifier(loss='exponential', max_depth=2)
    nn=
neural_network.MLPClassifier(hidden_layer_sizes=(nfeatsneural,nfeatsneural,nfeatsneural),
learning_rate_init=0.0001, max_iter=500)
    ab= ensemble.AdaBoostClassifier()
    bc= ensemble.BaggingClassifier(base_estimator=rf)
    vc= ensemble.VotingClassifier(estimators=[('gb', gb),('ab', ab),('bc', bc)], voting='soft')


    estimators= {'randomforest': rf,
                 'extratrees': et,
                 'kneighbors': kn,
                 'naivebayes': nb,
                 'decisiontree': dt,
                 'linearsvc': ls,
                 'gboost': gb,
                 'neuralnet': nn,
                 'adaboost': ab,
                 'bagging': bc,
                 'voting': vc
                 }

    train_results= {'fold':[], 'estimator':[], 'subjects':[],
                    'labels':[], 'predictions':[], 'scores':[],
                    'attempts':[]}
```

```
    test_results= {'fold':[], 'estimator':[], 'subjects':[],
                   'labels':[], 'predictions':[], 'scores':[],
                   'attempts':[]}

    for i in range(folds):
        train_ids=pd.DataFrame(index=outer_cv['train'][i])
        X_train= train_ids.join(data)
        y_train_df= train_ids.join(labels)
        y_train= np.array(y_train_df[y_train_df.columns[0]])

        test_ids=pd.DataFrame(index=outer_cv['test'][i])
        X_test= test_ids.join(data)
        y_test_df= test_ids.join(labels)
        y_test= np.array(y_test_df[y_test_df.columns[0]])

        for j,k in zip(estimators.keys(), estimators.values()):
            k.fit(X_train, y_train)

            predict_train= k.predict(X_train)
            train_scores= [1 if x==y else 0 for x,y in zip(y_train, predict_train)]
            train_results['fold'].extend([i+1]*len(X_train))
            train_results['estimator'].extend([j]*len(X_train))
            train_results['subjects'].extend(train_ids.index)
            train_results['labels'].extend(y_train)
            train_results['predictions'].extend(predict_train)
            train_results['scores'].extend(train_scores)
            train_results['attempts'].extend([1]*len(X_train))

            predict_test= k.predict(X_test)
            test_scores= [1 if x==y else 0 for x,y in zip(y_test, predict_test)]
            test_results['fold'].extend([i+1]*len(X_test))
            test_results['estimator'].extend([j]*len(X_test))
            test_results['subjects'].extend(test_ids.index)
            test_results['labels'].extend(y_test)
            test_results['predictions'].extend(predict_test)
            test_results['scores'].extend(test_scores)
            test_results['attempts'].extend([1]*len(X_test))

    train_df=pd.DataFrame.from_dict(train_results).set_index('subjects')
    test_df=pd.DataFrame.from_dict(test_results).set_index('subjects')

    train_df.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/combined-
study/outer_train_results.csv')
    test_df.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/combined-
study/outer_test_results.csv')
```

```
    print('TRAIN RESULT')
    trd= train_df.groupby('estimator').sum()
    trsum= (trd['scores']/trd['attempts'])*100
    print(trsum)
    trmax= trsum.idxmax(axis=1)
    print('\nBest train: {}\n'.format(trmax))

    print('TEST RESULT')
    ted= test_df.groupby('estimator').sum()
    tesum= (ted['scores']/ted['attempts'])*100
    print(tesum)
    temax= tesum.idxmax(axis=1)
    print('\nBest test: {}\n'.format(temax))

    return


def InnerFolds():
    a=input('Click and drag FEATURE SELECTED SINGLE FOLD DATA file here: ')
    a=a.strip('\' ')
    data=pd.read_csv(a, encoding='utf-8').set_index('PATIENT')

    b=input('Click and drag LABELS file here: ')
    b=b.strip('\' ')
    labels=pd.read_csv(b, encoding='utf-8').set_index('PATIENT')

    c=input('Click and drag SINGLE FOLD INNER CV file here: ')
    c=c.strip('\' ')
    with open(c, 'rb') as f: inner_cv= pickle.load(f)

    thisfold= input('Which fold is this? ')
    folds= len(inner_cv['train'])
    nfeatsmax= len(data.columns)
    nfeatsneural= round((nfeatsmax*2/3))

    rf= ensemble.RandomForestClassifier(max_features=nfeatsmax,
max_depth=5,bootstrap=False)
    et= ensemble.ExtraTreesClassifier(max_features=nfeatsmax, max_depth=5, bootstrap=False)
    kn= neighbors.KNeighborsClassifier(n_neighbors=nfeatsmax, p=1)
    nb= naive_bayes.GaussianNB()
    dt= tree.DecisionTreeClassifier(max_features=nfeatsmax, max_depth=5, criterion='entropy')
    ls= svm.LinearSVC(penalty='l1', dual=False)
    gb= ensemble.GradientBoostingClassifier(loss='exponential', max_depth=2)
    nn=
neural_network.MLPClassifier(hidden_layer_sizes=(nfeatsneural,nfeatsneural,nfeatsneural),
learning_rate_init=0.0001, max_iter=500)
```

```
ab= ensemble.AdaBoostClassifier()
bc= ensemble.BaggingClassifier(base_estimator=rf)
vc= ensemble.VotingClassifier(estimators=[('ab', ab),('gb', gb),('bc', bc)], voting='soft')

estimators= {'randomforest': rf,
            'extratrees': et,
            'kneighbors': kn,
            'naivebayes': nb,
            'decisiontree': dt,
            'linearsvc': ls,
            'gboost': gb,
            'neuralnet': nn,
            'adaboost': ab,
            'bagging': bc,
            'voting': vc
            }

train_results= {'fold':[], 'estimator':[], 'subjects':[],
                'labels':[], 'predictions':[], 'scores':[],
                'attempts':[]}

test_results= {'fold':[], 'estimator':[], 'subjects':[],
               'labels':[], 'predictions':[], 'scores':[],
               'attempts':[]}

for i in range(folds):
    train_ids=pd.DataFrame(index=inner_cv['train'][i])
    X_train= train_ids.join(data)
    y_train_df= train_ids.join(labels)
    y_train= np.array(y_train_df[y_train_df.columns[0]])

    test_ids=pd.DataFrame(index=inner_cv['test'][i])
    X_test= test_ids.join(data)
    y_test_df= test_ids.join(labels)
    y_test= np.array(y_test_df[y_test_df.columns[0]])

    for j,k in zip(estimators.keys(), estimators.values()):
        k.fit(X_train, y_train)

        predict_train= k.predict(X_train)
        train_scores= [1 if x==y else 0 for x,y in zip(y_train, predict_train)]
        train_results['fold'].extend([i+1]*len(X_train))
        train_results['estimator'].extend([j]*len(X_train))
        train_results['subjects'].extend(train_ids.index)
        train_results['labels'].extend(y_train)
        train_results['predictions'].extend(predict_train)
```

```
            train_results['scores'].extend(train_scores)
            train_results['attempts'].extend([1]*len(X_train))

            predict_test= k.predict(X_test)
            test_scores= [1 if x==y else 0 for x,y in zip(y_test, predict_test)]
            test_results['fold'].extend([i+1]*len(X_test))
            test_results['estimator'].extend([j]*len(X_test))
            test_results['subjects'].extend(test_ids.index)
            test_results['labels'].extend(y_test)
            test_results['predictions'].extend(predict_test)
            test_results['scores'].extend(test_scores)
            test_results['attempts'].extend([1]*len(X_test))

    train_df=pd.DataFrame.from_dict(train_results).set_index('subjects')
    test_df=pd.DataFrame.from_dict(test_results).set_index('subjects')

    train_df.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/combined-
study/inner_train_results_fold_'+str(thisfold)+'.csv')
    test_df.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/combined-
study/inner_test_results_fold_'+str(thisfold)+'.csv')


    print('TRAIN RESULT')
    trd= train_df.groupby('estimator').sum()
    trsum= (trd['scores']/trd['attempts'])*100
    print(trsum)
    trmax= trsum.idxmax(axis=1)
    print('\nBest train: {}\n'.format(trmax))

    print('TEST RESULT')
    ted= test_df.groupby('estimator').sum()
    tesum= (ted['scores']/ted['attempts'])*100
    print(tesum)
    temax= tesum.idxmax(axis=1)
    print('\nBest test: {}\n'.format(temax))

    return

def InnerHoldout():
    a=input('Click and drag FEATURE SELECTED SINGLE FOLD DATA file here: ')
    a=a.strip('\' ')
    data=pd.read_csv(a, encoding='utf-8').set_index('PATIENT')

    b=input('Click and drag LABELS file here: ')
    b=b.strip('\' ')
    labels=pd.read_csv(b, encoding='utf-8').set_index('PATIENT')
```

```
    c=input('Click and drag OUTER CV file here: ')
    c=c.strip('\' ')
    with open(c, 'rb') as f: inner_cv= pickle.load(f)

    thisfold= int(input('Which fold is this? '))

    nfeatsmax= len(data.columns)
    nfeatsneural= round((nfeatsmax*2/3))

    rf= ensemble.RandomForestClassifier(max_features=nfeatsmax,
max_depth=5,bootstrap=False)
    et= ensemble.ExtraTreesClassifier(max_features=nfeatsmax, max_depth=5, bootstrap=False)
    kn= neighbors.KNeighborsClassifier(n_neighbors=nfeatsmax, p=1)
    nb= naive_bayes.GaussianNB()
    dt= tree.DecisionTreeClassifier(max_features=nfeatsmax, max_depth=5, criterion='entropy')
    ls= svm.LinearSVC(penalty='l1', dual=False)
    gb= ensemble.GradientBoostingClassifier(loss='exponential', max_depth=2)
    nn=
neural_network.MLPClassifier(hidden_layer_sizes=(nfeatsneural,nfeatsneural,nfeatsneural),
learning_rate_init=0.0001, max_iter=500)
    ab= ensemble.AdaBoostClassifier()
    bc= ensemble.BaggingClassifier(base_estimator=rf)
    vc= ensemble.VotingClassifier(estimators=[('ab', ab),('gb', gb),('bc', bc)], voting='soft')

    estimators= {#'randomforest': rf,
                 #'extratrees': et,
                 #'kneighbors': kn,
                 #'naivebayes': nb,
                 #'decisiontree': dt,
                 'linearsvc': ls,
                 #'gboost': gb,
                 #'neuralnet': nn,
                 #'adaboost': ab,
                 #'bagging': bc,
                 #'voting': vc
                 }

    train_results= {'fold':[], 'estimator':[], 'subjects':[],
                    'labels':[], 'predictions':[], 'scores':[],
                    'attempts':[]}

    test_results= {'fold':[], 'estimator':[], 'subjects':[],
                   'labels':[], 'predictions':[], 'scores':[],
                   'attempts':[]}
```

```
    train_ids=pd.DataFrame(index=inner_cv['train'][thisfold-1])
    X_train= train_ids.join(data)
    y_train_df= train_ids.join(labels)
    y_train= np.array(y_train_df[y_train_df.columns[0]])

    test_ids=pd.DataFrame(index=inner_cv['test'][thisfold-1])
    X_test= test_ids.join(data)
    y_test_df= test_ids.join(labels)
    y_test= np.array(y_test_df[y_test_df.columns[0]])

    for j,k in zip(estimators.keys(), estimators.values()):
        k.fit(X_train, y_train)

        predict_train= k.predict(X_train)
        train_scores= [1 if x==y else 0 for x,y in zip(y_train, predict_train)]
        train_results['fold'].extend([thisfold]*len(X_train))
        train_results['estimator'].extend([j]*len(X_train))
        train_results['subjects'].extend(train_ids.index)
        train_results['labels'].extend(y_train)
        train_results['predictions'].extend(predict_train)
        train_results['scores'].extend(train_scores)
        train_results['attempts'].extend([1]*len(X_train))

        predict_test= k.predict(X_test)
        test_scores= [1 if x==y else 0 for x,y in zip(y_test, predict_test)]
        test_results['fold'].extend([thisfold]*len(X_test))
        test_results['estimator'].extend([j]*len(X_test))
        test_results['subjects'].extend(test_ids.index)
        test_results['labels'].extend(y_test)
        test_results['predictions'].extend(predict_test)
        test_results['scores'].extend(test_scores)
        test_results['attempts'].extend([1]*len(X_test))

    train_df=pd.DataFrame.from_dict(train_results).set_index('subjects')
    test_df=pd.DataFrame.from_dict(test_results).set_index('subjects')

    train_df.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/combined-
study/inner_holdout_train_results_fold_'+str(thisfold)+'.csv')
    test_df.to_csv(path_or_buf='/media/james/ext4data/current/projects/pfizer/combined-
study/inner_holdout_test_results_fold_'+str(thisfold)+'.csv')

    with open('/media/james/ext4data/current/projects/pfizer/combined-
study/trainedclassifier_innerfold_'+str(thisfold)+'.pickle', 'wb') as f: pickle.dump(k, f,
pickle.HIGHEST_PROTOCOL)

    print('D_-j RESULT')
```

```
    trd= train_df.groupby('estimator').sum()
    trsum= (trd['scores']/trd['attempts'])*100
    print(trsum)
    trmax= trsum.idxmax(axis=1)
    print('\nBest train: {}\n'.format(trmax))

    print('D_j (holdout for estimating model quality) RESULT')
    ted= test_df.groupby('estimator').sum()
    tesum= (ted['scores']/ted['attempts'])*100
    print(tesum)
    temax= tesum.idxmax(axis=1)
    print('\nBest test: {}\n'.format(temax))

    return
```

A1.6 Bootstrapping for Significance Testing

```
#!/usr/bin/env python3

import numpy as np
import pandas as pd
from collections import defaultdict
import matplotlib.pyplot as plt
import pprint, itertools, pickle, random, statistics

#Because we are sampling with replacement, we don't need to worry about the program picking
all subjects each time. Some may be picked more than once, and the total number of samples will
be equal to the number of subjects.

def Bill():
    a=input('Click and drag desired TEST RESULTS file (usually outer_test_results or
holdout_test_results: ')
    a=a.strip('\' ')
    otr=pd.read_csv(a).set_index('subjects')

    #Per subject accuracy
    acc= otr['scores']*100
    n= len(otr.index)
    runs= 10000
    chance=float(input('What % is chance? '))

    distribution= []
    for i in range(runs):
        sample= np.random.choice(acc, n,replace=True)
```

```
        sample_mean= sum(sample)/len(sample)
        distribution.append(sample_mean)

    dist_mean= sum(distribution)/len(distribution)
    p_value= sum(i<=chance for i in distribution)/runs

    print('{} runs, {} samples per run'.format(len(distribution), n))
    print('distribution mean: {}%'.format(dist_mean))
    print('p-value: {}'.format(p_value))

    bootstrap_results= {'samples per run': n,
                        'runs': 10000,
                        'distribution mean': dist_mean,
                        'p-value': p_value
                        }

    bdf= pd.DataFrame.from_dict(bootstrap_results, orient='index')

    binner=np.digitize(distribution, np.array(range(0,101)))
    plt.plot([chance,chance],[0,list(binner).count(statistics.mode(binner))],'-r',lw=2)
    plt.hist(distribution, bins=list(range(0,101)))
    plt.xlabel('% Accuracy')
    plt.ylabel('Number of runs')
    plt.title('Bootstrap sample distribution')
    plt.show()

    bdf.to_csv(path_or_buf='/media/james/ext4data1/current/projects/pfizer/refined-combined-
study/bootstrap_results.csv')

    return
```

A1.7 ROC Curve Creation

```
#!/usr/bin/env python3

import numpy as np
import pandas as pd
from collections import defaultdict
import matplotlib.pyplot as plt
import pprint, itertools, pickle, random, statistics
from sklearn import metrics

def Roc():
    a=input('Click and drag desired TEST RESULTS file (usually entire_dataset_results or
holdout_test_results: ')
```

```
    a=a.strip('\' ')
    results=pd.read_csv(a).set_index('subjects')

    labels= results['labels']
    predictions=results['predictions']

    fpr, tpr, thresholds= metrics.roc_curve(labels, predictions, pos_label=1)

    print('fpr: {} \ntpr: {}\n'.format(fpr[1]*100, tpr[1]*100))

    auc = "%.2f" % metrics.auc(fpr, tpr)
    title = 'ROC Curve, AUC = '+str(auc)
    with plt.style.context(('ggplot')):
        fig, ax = plt.subplots()
        ax.plot(fpr, tpr, 'darkorange', label='ROC curve')
        ax.plot([0, 1], [0, 1], 'k--', label='Baseline')
        plt.xlim([0.0, 1.01])
        plt.ylim([0.0, 1.01])
        plt.xlabel('False Positive Rate')
        plt.ylabel('True Positive Rate')
        plt.legend(loc='lower right')
        plt.title(title)
        plt.show()

    return

def HardPlace():

    return
```