Identification and Prediction of Incompletely Ascertainable, Rare Healthcare Outcomes

By

Alvin Dean Jeffery

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

December 18, 2021

Nashville, Tennessee

Approved:

Michael E. Matheny (Chair), MD, MS, MPH, FACMI

Ruth Reeves, PhD

Daniel Fabbri, PhD

ACKNOWLEDGEMENTS

We received support for this work from the Agency for Healthcare Research & Quality (AHRQ) and the Patient-Centered Outcomes Research Institute (PCORI) under Award Number K12 HS026395; resources and use of facilities at the Department of Veterans Affairs, Tennessee Valley Healthcare System, in collaboration with the Medical Informatics Fellowship; the Vanderbilt Institute for Clinical and Translational Research (VICTR) under Award Number UL1 TR000445 from NIH/NCATS; and the Advanced Computing Center for Research and Education (ACCRE) High-Memory Compute Nodes under Grant# 1S10OD023680-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of AHRQ, PCORI, the NIH, the Department of Veterans Affairs, or the United States Government.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1: Introduction

Clinical prediction models are increasingly common, particularly with advances in machine learning and artificial intelligence. A major bottleneck in the development and improvement of clinical prediction models is the assignment of an outcome label on which to train models. Traditionally considered the gold-standard for whether a patient experienced an outcome of interest, manual chart reviews are time-consuming and resource-intensive, particularly within extremely large data sets of patient records.

Our overall motivation in this work was to examine whether noisy labels generated from subject matter experts' heuristics using heterogenous data types could be used to provide labels to large, observational datasets to support predictive modeling.

Clinical Prediction Models

Clinicians have used current and historical patient data for diagnosis and treatment since the beginnings of medical care. For example, a patient presents with a set of symptoms or complaints that the clinician attempts to diagnose, usually with additional data from physical assessments and/or objective measures (e.g., laboratory studies). This diagnosis helps create a treatment plan that hopefully helps the patient return to their desired a state of health. A growth in biomedical knowledge (and technology) has enabled clinicians to use the same collected data not only to diagnosis a current problem but also to anticipate future problems a patient might have. A landmark example of this capability was the Framingham Heart Study in which approximately 5,000 individuals were prospectively monitored for several decades in order to evaluate their development of cardiovascular disease. Based on several risk factors (e.g., age, laboratory values), clinical prediction models have been developed to provide a new patient's risk of developing cardiovascular disease within as few as 6 years and as long as 30 years.[1,2]

Because it is infeasible to prospectively monitor thousands of individuals for every disease one could develop in the future, the number of predictive models remained relatively small until recently. Largely driven by the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 that provided financial incentives from the United States federal government for acute care hospitals to transition to EHRs,[3] the last two decades of healthcare informatics have brought widespread implementation of electronic health records (EHR).[4] The HITECH Act required hospitals to demonstrate "meaningful use" of EHRs through increasingly complex capture, use, and sharing of patient data in an electronic format.[5] The resulting emergence of extremely large datasets and accompanying growth in statistical processing capabilities have provided researchers and clinicians the ability to answer many questions that were not previously possible to answer. Many believe the use of clinical prediction models (also commonly referred to as predictive analytics) is the next step in expanding the clinician's toolkit because it provides a new set of information that can now be analyzed from available data.[6-8]

The purpose of predictive modeling is to collect and analyze data in real-time while providing end-users with a probability of a particular condition or event (e.g., hospital readmission, acute decompensation, or adverse drug events[6]). There are many published, peer-reviewed papers describing the performance of predictive analytic models in healthcare.[9,10]

Learning from Data

While predictive modeling continues to gain increasing attention in the healthcare community, the idea of creating a mathematical model to represent a phenomenon is not new. Researchers have developed models to learn from data for many years.[11,12] Such models can be used for activities beyond prediction, such as inference, including a focus on generating a probability distribution that a patient is associated with a particular condition in order to support phenotyping efforts.[13]

The three broadest approaches to modeling include: supervised, unsupervised, and reinforcement learning. While all modeling approaches require numerical representations of input measures (or *features*) created from source data, each approach differs with respect to the outcome measure. Supervised models require an outcome measure and have the goal of predicting the value of that measure based on the inputs. Unsupervised methods have no outcome measure but rather identify the patterns or groupings within the inputs.[12] Reinforcement learning methods use a reward system in which an agent seeks to control a system, using a numerical reward that the agent seeks to maximize as it moves through each state.[14] Supervised models are the most widely used within healthcare and are the primary focus here.

Common supervised learning techniques include regression, linear discriminant analysis, k-nearest neighbors, decision trees, random forests, Naïve Bayes, and multilayer perceptrons (i.e., neural network).[12,15,16] Regression modeling is a foundational technique that has been one of the longest-used methods. The use of regression methods benefits from familiarity by many investigators from many disciplines as well as easier interpretability of the results. A disadvantage of regression is the requirement to meet many statistical assumptions of the data (compared to other methods) and the need to explicitly specify interactions between input variables. Linear discriminant analyses are used to separate the data into maximally-different groups but requires several assumptions of the underlying data, including that all input measures be on the continuous scale. The k-nearest neighbor algorithm uses similar records in the data to label the new record based on a majority vote and extends well to multi-class labels; however, there is not clear guidance on the appropriate value of how many *k* neighbors should be specified maximize performance. The Naïve Bayes algorithm is very simple to implement because it has no hyper-parameters to specify; however, it does not allow for interactions between variables and assumes complete independence of the classes. Random forests aggregate many decision trees by splitting input variables at a threshold that maximizes the

difference between classes. Random forests require few assumptions of the input data and handle large data and variable interactions easily. Multilayer perceptrons are the newest algorithms to be widely used. Multilayer perceptrons have the benefit of very few assumptions about the underlying data but require extremely large data sets, are difficult to interpret *how* the models work, and are computationally intensive.

Several quantitative metrics exist to evaluate how well these supervised methods can predict an outcome. Common metrics include: (a) sensitivity – the proportion of records that truly have the condition that are labeled by the model as having the condition, (b) specificity – the proportion of records that truly lack the condition that are labeled by the model as not having the condition, (c) positive predictive value – the proportion of records that are labeled as by the model as having the condition that truly have the condition, (d) F1 score – the harmonic mean of sensitivity and positive predictive value, (e) area under the curve (AUC) – a summary statistic the combines the sensitivity and specificity of the model over all threshold values, (f) accuracy – the proportion of correct predictions, and (g) mean squared error – the square root of the sum of differences between predicted and actual values after being squared and divided by the number of records.[12,17]

Challenges in Developing Clinical Prediction Models

Although data availability and scientific/statistical computing capacity have increased, developing high-performing clinical prediction models remains a challenge due to the complexity of attempting to model phenomena in the natural world (including healthcare). Common challenges include, but are not limited to: (a) selecting a statistical or machine learning method that meets the assumptions of predictor and outcome variables, (b) dealing with missing data, (c) variable selection, dimensionality reduction, and feature engineering, and (d) validation techniques.[11,18,19] Addressing these challenges has been the focus of many statisticians, data scientists, informaticians, and others for several decades. There are many proposed solutions to these challenges, but they are beyond the scope of this thesis work.

An additional challenge that has received less attention when working with larger data sets is generating reliable outcome labels in large enough quantities to function as training data. Manual (human) review of records has traditionally been considered the gold-standard of phenotyping. Unfortunately, manual reviews are time-consuming and resource-intensive, particularly within extremely large data sets of patient records.[20,21] Therefore, **the primary focus of this work was the development of a phenotyping process that assigns outcomes labels to large data sets for downstream tasks, such as predictive modeling.**

Phenotyping within Large Data Sets

To overcome the challenge of manual reviews within EHRs, two broad approaches have been used: (a) condition-specific algorithms that leverage rule-based logic incorporating diverse data sources such as diagnostic billing codes, clinical notes, and laboratory values, among others,[22,23] and (b) high-throughput methods that assign thousands of phenotypes to the EHR data, such as PheCodes which are groupings of diagnostic billing codes.[23-26] Condition-specific algorithms can be iteratively developed to improve recall and precision. High-throughput methods have the benefit of greater generalizability across institutions as well as the speed and scale by which they can apply numerous phenotypes. However, drawbacks of high-throughput methods include: (a) the available phenotypes might not include the specific phenotype identified *a priori* by an investigator, and (b) investigators are sometimes more interested in less well-defined phenotypes that do not have diagnostic billing codes (e.g., adverse events) or where coding practices change for policy reasons (e.g., opioid use disorder-related diagnoses).

Both unsupervised and supervised methods have been examined for their ability to assist with phenotyping.[22] In unsupervised methods, features extracted from source data are engineered with the goal of organizing records into similar clusters, topics, or sub-types that are described or characterized by subject matter experts,[22,27] including the potential to describe clusters as containing either cases or controls.[28] While unsupervised methods can be helpful for

identifying latent groupings within a population (e.g., more granular sub-types of diabetes), they do not facilitate the application of a phenotype an investigator seeks to identify *a priori*.

In supervised methods, features extracted from source data are engineered to serve as predictors in statistical or machine learning models where an outcome is supplied in at least some of the data. Therefore, the use of supervised algorithms for phenotyping is primarily helpful when one has a training set in which labels have already been applied and the goal is to conduct phenotyping with the same input features in a new data set. The task of assigning outcome labels to each record within at least one training set remains.

To expedite the process of labeling within larger data sets, some have advocated for the use of *noisy* labels. When using noisy labels, investigators make the assumption that a large set of data with imperfect labels (i.e., some inaccuracies present) could train a machine learning model just as well as smaller data sets with clean (i.e., high confidence in accuracy) labels.[13,29,30] Noisy labels can be derived from processes such as using experts to create a set of heuristics that generate an approximation of the ground truth label across all records or using a readily-available proxy label that is correlated with the ground truth label. A promising starting point that balances the knowledge of subject-matter experts with the speed of data-driven approaches is the use of *anchor learning*.[31] In anchor learning, a subject matter expert creates a set of imperfect rules that satisfy two conditions: (1) the rule has high positive predictive value and (2) conditional independence wherein no additional information would facilitate improved labeling if the label were already known.[31] These rules serve as an imperfect (or *noisy*) label on which to build supervised models that can generalize beyond the specified anchors and yield the probability of a record having the label of interest. This framework has been applied to phenotyping for healthcare and standardized within the Observational Health Sciences and Informatics (OHDSI) network and is known as the Automated Phenotype Routine for Observational Definition, Identification, Training, and Evaluation (APHRODITE).[32]

A related, newer, and less-evaluated framework is the specification of multiple imperfect

labels developed by subject matter experts. In this **data programming paradigm** (introduced by Ratner et al. at Stanford University[33] and incorporated into a software program known as Snorkel), a developer writes multiple *labeling functions* (LF) that serve as noisy labels based on heuristics, patterns, or external information. Each LF processes input data and returns a vote of a Yes (1), No (0), and/or Abstain (-1). LFs can be overlapping such that multiple LFs might use the same input data. LFs can be conflicting such that the same record yields different votes (e.g., one LF yields a Yes vote while another LF yields No vote on the same record). While LFs could potentially return any of the three vote options, each LF only needs to return 2 of the 3 vote options (e.g., Yes versus Abstain, Yes versus No). LFs produce an *m x n* label matrix with *m* examples and *n* LFs. Without any ground-truth data, Snorkel uses the label matrix to model accuracies and correlations between LFs to optimize a Generative model that yields probabilistic labels. Then, probabilistic labels are used to train a Discriminative model with any statistical or machine learning model of the developer's choice.

While the Discriminative model is not technically necessary, this step confers the added benefit of increased generalizability. A Discriminative model can be used by external stakeholders or with future unlabeled data without needing the Snorkel software, LFs, or access to the same input data used for the Generative model. For example, a single LF for a Generative model might comprise complex if-else logic related to the number of clinical notes written by respiratory therapists and whether the patient has an extended mechanical ventilation period and certain respiratory conditions. This complex logic yields a simple vote of Yes, No, or Abstain for populating a column in the label matrix. However, a Discriminative model could include separate features for respiratory therapy clinical note counts, actual mechanical ventilation duration, and a binary indicator of existing respiratory conditions. Each of these features would serve as a separate predictor in the Discriminative model. Additional benefits of using a Discriminative model are: (a) the ability to include the labels from manually-reviewed records (e.g., those used during development and validation of the Generative model) as labels

to improve the model's performance as well as (b) the option to use a noise-aware Discriminative model that can account for the uncertainty within the probabilistic Generative labels (e.g., models that can use probabilities as outcome labels or models that allow sample weighting during the training process).

As with any modelling approach, one can include features engineered from a variety of sources when developing either the Generative or Discriminative models. Structured data, as previously described in the condition-specific and high-throughput methods, are commonly leveraged. Unstructured data (e.g., text-based notes) can also be included after they undergo natural language processing (NLP) techniques in order to be transformed into computable features. NLP methods for modelling may include, but are not limited to: bag-of-words, keyword searching, concept extraction,[22] and vector embeddings.[34] NLP methods tend to perform better in highly-specific domains rather than being generalized techniques that apply to a multiple clinical domains and phenotypes.[35]

A benefit of the data programming approach over techniques such as the semi-supervised PheCAP[26] is that phenotypes need not be well-established, universally agreed-upon phenotypes. While these phenotypes will not be perfect, it is possible to specify the amount of uncertainty in the estimates. Describing the uncertainty allows each user (i.e., research investigator, policy maker) to make decisions on whether the phenotype can be used for their purposes.[36]

Use Case: Opioid-Induced Respiratory Depression

In our exploration of predictive modeling approaches, we selected opioid-induced respiratory depression (OIRD) as our clinical use case. With almost 1 in 10 hospitalized patients experiencing an adverse event annually,[37] improving patient safety is greatly needed. Among perioperative patients, respiratory failure is the most common adverse event (9.13 per 1000 patients[38]) and costs up to $23.5 billion annually.[39] Respiratory depression can occur in a variety

of hospitalized patients, but surgical patients are particularly susceptible due to opioid administration for postoperative analgesia. While attention to opioid problems has increased in recent years,[40-42] adverse events associated with opioid administration are not new to healthcare. A 2004 literature review of 165 papers revealed an OIRD incidence of 0.1-1.3% with a definition of naloxone administration, 0.7-1.7% with a definition of hypoventilation, 1.4-7.6% with a definition of hypercarbia, and 10.2-26.9% with a definition of oxygen saturation.[43] A more recent (2018) literature review of 13 studies reported a similar incidence with a total average of 0.5%.[44]

Predictive Models for OIRD

Operational definitions of OIRD have included naloxone administration, hypoventilation, hypercarbia, and oxygen de-saturation.[43,45] The lack of a standardized definition makes building predictive models (and comparing performance between different models) challenging due to the difficulties in the assignment of outcome labels, both in the setting of manual chart reviews and automated approaches. In manual chart reviews, it can be difficult for a reviewer to determine if naloxone administration resulted in the intended benefit. It is not uncommon for naloxone to be administered in the setting of altered mental status as a way of determining if opioids are responsible. However, simply because a patient is receiving opioids does not mean that is the etiology of their altered mental status. Similarly, a patient could experience hypoventilation or hypoxia as a result of a non-opioid related disease process.

To our knowledge, there is no automated approach to identifying OIRD. The most relevant criteria for identifying patients with OIRD at scale would be Patient Safety Indicator (PSI) 11 focused on post-operative respiratory failure from the Agency for Healthcare Research and Quality (AHRQ).[46,47] Although the criteria focus on respiratory failure, they are not specific to opioids. External validation studies of the PSI-11 criteria have demonstrated a sensitivity of 0.19-0.44 and a positive predictive value of 0.4-0.83 for post-operative respiratory failure.[48-52] However, notably in one study,[48] oversedation was not the cause of any respiratory failure

events. Therefore, how well PSI-11 criteria identify OIRD is unknown.

Regarding predictors, adverse events are commonly attributed to organizational or systems factors,[37] but biological factors also play a role in the respiratory failure cases associated with opioid administration.[44] In a 2018 literature review of OIRD risk factors comprising 13 articles, Gupta et al.[44] identified several possible risk factors. Surgical risk factors included: first 24 hours after surgery, orthopedic and transplant surgeries, greater than 60 years of age, female gender, American Society of Anesthesiologists' Physical Status Classes 3 and 4, opioid dependence, and genetic polymorphisms. Comorbidities included: diagnosed or suspected obstructive sleep apnea; many cardiovascular diseases; diabetes mellitus; obesity; and renal, pulmonary, neurological, and liver diseases. Peri-operative risk factors included: respiratory events in the post-anesthesia care unit; concomitant sedative use; patient-controlled analgesia administration; excessive opioids; multiple routes of opioid administration; multiple prescribers; two or more opioids; excessive sedation; inadequate monitoring; hyperoxemia; and supplemental oxygenation. These risk factors could serve as a candidate features for a clinical prediction model. At the beginning of this thesis work, we found no OIRD clinical prediction models in the literature.

To build a model, it is important to include predictors that would best help predict the outcome, drawing on these data streams: structured data, unstructured data, and genetic data. Predictors derived from structured and unstructured data within the EHR have been used extensively within clinical prediction models, but genetic data are not routinely included.

Several studies have explored the influence of genetic variants in OIRD. If simple genetic etiologies of OIRD can be identified and used to decrease its incidence, we have the opportunity to provide individualized treatments (e.g., changing opioid type or dose) to mitigate adverse events while reducing overall healthcare costs. Previous studies have identified approximately 6 statistically significant single nucleotide polymorphisms (SNPs) associated with OIRD, but the small sample sizes (largest n=347) and suboptimal analysis methods leave

several gaps for continued exploration. Given that SNP identification costs less than $2,[53-55] collecting relevant SNP data on the 16 million U.S. surgical inpatients every year[56] could be a cost effective approach to OIRD reduction.

Genetic Associations with OIRD

Given the current understanding of opioid pathways, which include numerous proteins present in cellular membrane opioid receptor sites, intra-cellular pathways, and liver metabolism, many genes could influence an organism's response to opioids. In older studies of knockout mice, researchers discovered a lack of respiratory depression when administering morphine to mice with the μ receptor removed,[57] and there was less respiratory depression in those with μ receptor deficiencies following morphine-6-glucuronide (morphine's active metabolite) administration.[58] One study of 87 human brain autopsy tissue samples suggested the 118A>G variant significantly reduces mRNA and μ receptor protein production.[59] Some researchers have noted that human cell lines with the OPRM 118A>G variant have μ receptors with lower binding-site availability;[60] however, another study of the same cell line (i.e., HEK293) did not find evidence of reduced receptor function.[61] Mura et al. explicitly noted the inability to extrapolate findings to clinical scenarios due to current conflicting basic science evidence and additional complexities of the clinical environment.[62] The most commonly explored genetic variant (likely due to its high frequency in many populations) is SNP 118A>G (rs1799971) found on the μ-opioid receptor gene OPRM1. The 118A>G variant ranges in frequency from 0.8% (Sub-Saharan ethnicity) to 8.2-17% (Caucasians) to 48.9% (Asians).[63] Other common variants are those found in the CYP2D6, ABCB1, UGT, and COMT genes. Table 1 lists most of the variants that have been proposed as having an association with opioid effects. The following sections describe published reports of genetic variations associated with opioid-induced respiratory depression in clinical settings with human subjects.

Case Reports: Although they provide low-level evidence for decision-making, case reports can help generate hypotheses for developing larger studies. One report of a fatal hydrocodone overdose in a child identified poor first-phase metabolism as a result of a functionally impaired CYP2D6 (with a *2A/*41 variant).[64] Another report by the same authors described two adult patients who received morphine in the perioperative setting and subsequently developed respiratory depression. The first patient had both a UGT2B7 C802 TT genotype (rs7439366) variant, which might have increased opioid metabolite formation, and a COMT haplotype TCA (rs4633, rs4818, rs4680), which might have increased opioid sensitivity. The second patient had an ABCB1 haplotype TTT (rs1128503, rs2032582, rs1045642) and COMT haplotype CCG (rs4633, rs4818, rs4680), which might have increased opioid sensitivity, and conversely, an OPRM1 variant (rs1799971 [G/G]), which might increase opioid requirements.[65] Finally, a patient receiving tramadol who experienced respiratory depression identified a CYP2D6 UM variant (i.e., an “ultra-metabolizer” with a duplicate gene more susceptible to increases in active metabolite concentrations). The author also reported 3 other similar cases of codeine- or tramadol-induced respiratory depression in patients with CYP2D6 duplications.[66] Similar case reports of codeine- and morphine-related respiratory depression have been noted among patients with CYP2D6 polymorphisms, and the Food and Drug Administration now how several warnings related to codeine.[67-70]

Table 1.1. Genetic variants with proposed opioid effects.

Gene	SNP Location (rs…)	Potential Effect(s)[68,71-74]
ABCB1	1128503, 2032582, 1045642	Decreased intestinal expression,[71] Reduced function,[68] and Respiratory depression[73]
	9282564	Reduced function[68] or Respiratory depression[73]
	2229109	Respiratory depression[73]
ADRB2	11958940, 1432623, 2400707, 1042713, 1042714, 1042717	Unspecified[72]
ANKK1	1800497	Unspecified[72]
COMT	4633, 4818, 4680	Decreased opioid response[71] and Reduced function[68]
	6269	Unspecified[72]
FAAH	324420, 932816, 4141964, 3766246, 324419, 2295632	Enhanced response[75]
CYP2D6	16947, 1135840, 769258	Rapid metabolism[71] or Reduced function[68]
	28371706, 28371725	Reduced function[68]
	35742686, 3892097	Loss of function[68]
	5030655	Poor metabolism[71] and Loss of function[68]
	5030867, 1065852, 1065858	Loss of function[68]
CYP2B6	2279343, 34223104 3211371, 3745274, 2279343, 28399499	Enhanced function[68] Reduction function[68]
CYP3A4	2740574 35599367	Unknown[71] Reduced function[68]
CYP3A5	776746 10264272, 41303343	Poor metabolism[71] and Loss of function[68] Loss of function[68]
DRD2	2734838, 6279	Unspecified[72]
GCH1	4411417, 3783641, 8007267, 752688	Unspecified[72]
MC1R	1805005	Loss of function[71]
MRP2	n/a	Impaired activity[71]
OATP2	4149056	Possible transporter for endogenous opioids[71]
OPRM1	1799971	Higher morphine requirement[71] and Impaired receptor function[68]
	1799974	Altered receptor signaling[71]
	2234918	Unspecified[72]
	Methylation	Decreased expression[74]
OPRK1	1051660, 702764	Unknown[71]
OPRD1	1042114, 2234918	Unknown[71]
SLCO1A2	11568563	Unknown[71]
SLCO1B3	4149117, 7311358	Unknown[71]

TRPA1	222747, 13279503, 1947913, 13255063, 3735942, 3735943, 1443952, 1025928, 1198795	Unspecified[72]
UGT1A1	35350960, 815347	Decreased expression[71] and Reduced function[68]
	8175347	Enhanced function[68]
UGT2B7	7439366, 12233719, 7438135	No differences[71] or Reduced function[68]
	7668258	Unspecified[72]

Prospective Cohort Studies*:* A group of researchers at Cincinnati Children's Hospital Medical Center has been particularly productive in publishing their prospective cohorts involving pediatric surgical patients treated with morphine. Their studies with a clinical outcome comprising at least OIRD include a GWAS with 259 children;[75] a regression analysis focused on a few SNPs with 88 children[76] and 263 children;[73] a single SNP with 101 children;[77] and a cluster analysis with 347 children.[72] With the exception of Biesiada et al. who explored a panel of 42 SNPs,[72] these researchers have primarily focused on genetic variants within ABCB1, FAAH, and OPRM1 genes. The studies were well-described and attempted to adjust for potential confounders; however, in my opinion, they did not use the best regression modeling approaches to minimize false discovery in small samples while including all relevant covariates. Effect sizes from these studies suggest children with genetic variants could have up to 4.7 times greater odds (95%CI: 2.1-10.8)[73] or 2.1-3.8 times greater relative risk [72] of OIRD compared to wild-type genotypes.

Exploring fentanyl administration in the perioperative setting with Korean adult patients, the presence of a TTT haplotype (at rs1128503, rs2032582, rs1045642) on the ABCB1 gene increased the risk of OIRD.[78] Conversely, no associations were found with OPRM1 118A>G and respiratory depression in surgical Han Chinese adult patients.[79] The study's analyses included only simple tests of differences (i.e., ANOVA, t-test) without adjusting for covariates. Henker et al. found the 118A>G variant to be associated with less sedation (albeit, with a very small effect size and no subsequent respiratory depression) in a sample of 79 patients after adjusting for

several covariates in a multivariable regression analysis. However, they applied questionable exclusion of patients and variables in the analysis.[80]

After my academic work began, a relevant prospective clinical trial was registered with clinicialtrials.gov (NCT03441281). The researchers aim to examine the influence of several (unspecified) candidate SNPs on fentanyl-induced OIRD in the pre-operative setting before traditional anesthesia induction agents are administered. At the writing of this thesis (Fall 2021), no results were available.

Experimental Studies: To our knowledge, the only randomized trial exploring the association between the OPRM1 118A>G variant and OIRD was a small sample of 16 healthy adults randomized to 4 different groups of morphine-6-glucuronide (morphine's active metabolite) administration in a laboratory environment. The study found no association with OIRD even though pain levels were higher in those with the variant.[81] A similar study with 20 healthy adults who were selected to represent all 3 possible 118 variants (AA, AG, and GG) found higher alfentanil doses were needed in the presence of a G variant with respiratory depression similar in the AA vs. AG groups and no respiratory depression in the GG group.[82] Another experimental study with a sample of 33 healthy adults participated in a cross-over study where they were randomized to an oxycodone or placebo arm and then moved to the opposite study arm after a 1-week washout period. The study focused on pain responses and a few adverse events, and the authors reported a number of variants that could play a role in pain response and adverse events.[83]

Summary of Genetic Influence on OIRD: Non-clinical studies in mouse and human models suggest the μ receptor is responsible for OIRD given that deficiency or absence of μ receptors curtail OIRD. A variety of case reports and cohort studies have identified associations between genetic variants (particularly CPY2D6) and OIRD. The cohort studies had relatively small sample sizes and primarily included perioperative patients. Further, results from cohort trials revealed some contradictory findings, and among the cohort trials with statistically

significant associations, the effect sizes were typically small. I also question several of the authors' analytical choices for developing their regression models (e.g., variable selection based on association with the outcome, assuming linearity of covariates, and excluding patients for any missing outcome). In spite of several studies providing evidence of reduced opioid effects in carriers of a G variant on OPRM1 118,[84] a 2009 meta-analysis of OPRM1 118A>G variants included 5 studies where opioid side effects were considered as outcomes (sample size not reported but less than 1,480), and no significant association was found with respiratory depression.[63] Given the reported statistically significant findings from my literature review, Table 2 identifies the genetic variants possibly associated with OIRD.

Gaps in Knowledge of Genetic Influence on OIRD: While the previously described works identify the role of genetic variants in OIRD with 8 of the ~80 hypothesized SNPs yielding statistically significant associations, the small sample sizes and suboptimal analysis methods leave several gaps for continued exploration. The current state of OIRD genetic association evidence does not warrant modifying opioid administration clinical decisions based on an individual patient's genetic profile. A literature review made a similar conclusion (with the exception of codeine and tramadol administration in the presence of CYP2D6 polymorphisms as previously described).[68]

Table 1.2. Possible associations between genetic variants and opioid-induced respiratory depression.

Gene	Variant	Opioid	Supporting Evidence*
ABCB1	TTT (rs1128503, rs2032582, rs1045642)	Morphine	Case Report[65] Prospective trial (n=347) of peri-operative children[72] [rs1045642 only]
		Fentanyl	Prospective trial (n=126) of peri-operative adults[78]
	GG, GA (rs9282564)	Morphine	Prospective trial (n=263) of peri-operative children[73]
ADRB2	rs1042713	Morphine	Prospective trial (n=347) of peri-operative children[72]
COMT	TCA & CCG (rs4633, rs4818, rs4680)	Morphine	Case Report[65]
FAAH	rs324420	Morphine	Prospective trials (n=259 and n=101) of peri-operative children[75]
	rs2295632	Morphine	Prospective trial (n=347) of peri-operative children[72]
CYP	2D6 *2A/*41	Hydrocodone	Case Report[64]
	2D6 UM (Duplication)	Tramadol	Several Case Reports[66]
OPRM1	118A>G (rs1799971)	Morphine	Case Report[65] Prospective trial (n=88) of peri-operative children[76] Experimental study (n=16)[81]
		Morphine-6-glucuronide Alfentanil	Experimental study (n=20)[82]
UGT2B7	C802 TT (rs7439366)	Morphine	Case Report[65]

**Note:* All variants listed for case reports. Only statistically significant variants listed for experimental and quasi-experimental trials.

For case reports, there were many confounders (e.g., other medications metabolized in the same pathway as opioids, renal dysfunction) that might also contribute to respiratory depression. Most cohort studies exploring opioid pharmacogenetics have focused on pain control and dosage requirements as outcomes with far fewer including OIRD as an outcome. Regression analyses routinely included several covariates (e.g., age, sex, race, morphine dosage, pain scores); however, they removed variables that did not demonstrate strong predictive influence of OIRD before including genetic variants, a method which can produce biased results.[11] Similarly, no study included all potential genetic variants that align with the (currently understood) opioid pathway even though attention to gene-gene interactions for OIRD

has increased since approximately 2009.[85] For example, even though OPRM1 118A>G variants might reduce OIRD (due to reduced receptor activity), a CYP2D6 duplication could increase OIRD (due to rapid metabolism to more potent metabolites). Even if we could control for interactions in genetic variations, there are many non-genetic factors that also moderate respiratory depression, such as concurrent medications, renal function for drug secretion, cognitive status, and possibly epigenetic changes.[86] Finally, several studies used vitals in the immediate post-operative period as indicators of OIRD; however, this approach can greatly limit sample size and misses many patients who develop OIRD after returning to an inpatient hospital bed.

Summary

In summary, OIRD is an important clinical condition that affects a non-trivial number of patients receiving opioids. A clinical prediction model to identify those at highest risk for OIRD could be beneficial for prevention efforts. While all clinical prediction models have challenges in their development and evaluation, an OIRD model is further complicated by the challenge of ascertaining which patients experienced OIRD and thus assigning the appropriate outcome label for modeling. This thesis work addressed two main challenges in developing such a prediction model: (1) assigning outcome labels to a large observational data set without relying on manual chart reviews and (2) including heterogenous data types as predictors. If a clinical prediction model could be used to decrease OIRD incidence, we have the opportunity to provide individualized treatments to mitigate adverse events while reducing overall healthcare costs.

Chapter 2: Use of Noisy Labels as Weak Learners to Identify Incompletely Ascertainable Outcomes: A Feasibility Study with Opioid-Induced Respiratory Depression

Introduction

Before a prediction model can be created to identify high-risk patients, we first need a large dataset of labeled data on which to train a model. Our specific aim in this study was to use an ensemble of clinically-informed noisy labels that act as weak learners to create outcome labels for post-operative opioid-induced respiratory depression (OIRD) in a large, observational data set.

Methods

Design

We applied the *data programming* paradigm known as Snorkel, which we described in Chapter 1.

Sample & Setting

We collected data from post-operative adult patients in the de-identified electronic health record at Vanderbilt University Medical Center making use of the BioVU Sample Repository and "Synthetic Derivative" databases. The Synthetic Derivative (SD) is a de-identified copy of the main hospital medical record databases created for research purposes. The de-identification of SD records was achieved primarily through the application of a commercial electronic program, which was applied and assessed for acceptable effectiveness in scrubbing identifiers. For instance, if the name "John Smith" appeared in the original medical record, its corresponding record in the SD does not contain "John Smith". Instead, it is permanently replaced with a tag [NAMEAAA, BBB] to maintain the semantic integrity of the text. Similarly, dates, such as "January 1, 2004" have been replaced with a randomly generated date, such as "February 3,

2003."

We limited the cohort to surgical procedures eligible for inclusion in the Agency for Healthcare Research & Quality's (AHRQ) Patient Safety Indicator-11 (Postoperative Respiratory Failure Rate).[46,47] These criteria exclude procedures with increased risk for respiratory failure (e.g., airway/lung & esophageal procedures) as well as people with degenerative neurological disorders. The AHRQ criteria also restrict encounters to elective surgical procedures, which is a data element not available in our de-identified database. As a proxy for elective status, we chose to exclude encounters where the qualifying surgical procedure occurred on the same day as an Emergency Department visit.

Our cohort comprised 52,861 visits representing 44,999 patients, which we divided into Training, Development, Validation, and Testing sets (see Table 1). We initially created the Test Set based on those with available genetic data (n=2,189 patients). We included all visits that met AHRQ PSI-11 criteria (n=264, 0.50% of cohort) and randomly sampled 500 visits (0.95% of cohort) that did not meet AHRQ PSI-11 criteria. Of the remaining 52,097 visits, we excluded 285 of those visits from further assignment because they were associated with patients who had visits already included in the Test Set. Then, we randomly selected 50 visits for the Validation Set and Development Set using 2:1 over-sampling based on AHRQ criteria with 2 AHRQ-defined cases per 1 AHRQ-defined control.

Table 2.1. Characteristics of data sub-sets for study.

Data Set	**Sample Size**	**Purpose**	**Selection Process**
Test	764	Final evaluation of Discriminative model	Random from those with genetic data
Validation	90 (originally 50)	Discriminative model selection	Random with oversampling from AHRQ criteria
Development	90 (originally 50)	LF development & Generative model validation	Random with oversampling from AHRQ criteria
Training	51,632 (originally 51,712)	Generative model development	Not in Test, Validation, or Development sets

Generative Model Development & Evaluation

Developing the Generative model involved an iterative process of: (a) developing candidate LFs, (b) examining performance of candidate LFs in the Development Set, (c) using the Snorkel paradigm to develop a candidate Generative model in the Training Set, and (d) evaluating performance of the candidate Generative model in the Development Set.

The dually trained biomedical informaticist and critical care nurse (ADJ) conducted chart reviews of visits in the Development Set to create candidate LFs in Python and determine whether each visit had evidence of OIRD. LFs comprised data from medication information, clinical note text (using regular expressions for words and short phrases), and administrative diagnostic and procedure codes. In contrast to Snorkel's recommended context hierarchy, we found collapsing all relevant data for a visit into a single row for LF application improved performance. Performance metrics guiding LF creation and modification included: (a) coverage – the proportion of visits in which the LF could yield a vote, (b) conflicts – whether another rule yielded a different vote, and (c) empirical accuracy – the proportion of visits correctly labeled, excluding Abstain votes, based on the single reviewer's determination. We used the performance metrics to iteratively modify LFs.

Following LF creation and modification, we used Snorkel's paradigm to generate a probability of whether a visit included an OIRD event. We conducted hyper-parameter tuning of the neural networks using learned LF weights in the Training Set and empirical accuracy in the Development Set (except in the final round where we combined the Training Set and Development Set and used Validation Set to assess empirical accuracy). We selected hyper-parameters that, in general, yielded higher LF weights for clinically-important rules. For example, an LF that uses information about naloxone administration (i.e., a specific treatment for OIRD reversal) should be more important than an LF that assess for altered mental status, which is less specific to OIRD.

Due to the low number of positive Cases in the initial Development Set (2/50, 4%), we

applied this iterative process to enrich the Development Set and Validation Set by extracting visits with the top 20 probability values from the Training Set and dividing those equally among the Development Set and Validation Set. After Round 4, the primary LF developer facilitated a focus group with clinicians and biomedical informaticists to discuss face validity of the current LFs and solicit additional heuristics for additional LFs. Table 2 delineates the staged process of iterative LF development through 5 rounds.

Table 2.2. Labeling function (LF) development process with Training and Development Sets.

Round	**Visits with OIRD**	**Post-Review Actions**
1. Review 50 Development Set visits	2/50	-Draft LFs. -Extract top 20 from Training Set, sending 10 to Development Set & 10 to Validation Set
2. Review 10 new Development Set visits	6/10	-Modify LFs & Add LFs -Repeat top 20 extraction
3. Review 10 new Development Set visits	1/10	-Modify LFs & Add LFs to correct for overfitting -Repeat top 20 extraction
4. Review 10 new Development Set visits	9/10	-Solicit feedback from clinicians & biomedical informaticists on LFs -Modify LFs & Add LFs
5. Review 10 new Development Set visits	9/10	-Create final Generative model in the combined Training & Development Sets

Discriminative Model Development & Evaluation

We used the final Generative model's probabilistic labels as the outcome labels for developing a Discriminative model. Unlike the Generative model that makes predictions using the output from LFs, the Discriminative model uses features directly from the source data. In our model, we selected age, gender, binary indicators related to administrative codes and naloxone administration, and frequency of keywords/phrases in clinical notes to serves as predictors. Administrative codes included diagnostic and procedure codes related to respiratory failure/disease, prolonged mechanical ventilation, sepsis, cardiovascular disease, and cerebrovascular accidents. Keywords and phrases related to naloxone administration and its effectiveness, narcotic overdose, absence of pain medications, decreasing or holding opioids, presence of acute events, altered mental status, pinpoint pupils, and hypoxia. We also included

the number of notes from respiratory therapists and mentions of rapid response teams.

We began Discriminative model development with off-the-shelf[12] machine learning algorithms from Python's scikit-learn to identify the most promising algorithms for hyper-parameter tuning. Classification algorithms comprised logistic regression, linear discriminant analysis, k-nearest neighbors, decision trees, random forest, Naïve Bayes, and a multilayer perceptron (i.e., neural network).[12] Regression algorithms comprised linear regression, random forest, and a multilayer perceptron.[12] Based on F1-scores, AUC, and mean squared error,[17,19] we chose the random forest and multilayer perceptron algorithms for hyper-parameter tuning[12] in both the classification and regression tasks.

To estimate the Discriminative model's future performance in an unbiased manner, we performed nested cross-validation with a manual grid search on the combined Training/Development Set using 3 inner folds and 10 outer folds. The nested cross-validation suggested F1 scores will range 0.6-0.8 for classifiers and 0.4-0.7 for regressors, AUCs will range 0.75-0.9 for classifiers and 0.6-0.8 for regressors, and mean squared errors will range 0.005-0.008 for classifiers and 0.005-0.01 for regressors (see Figures 1-3).

Figure 2.1. Comparison of F1 score values from nested cross-validation of the Discriminative model.

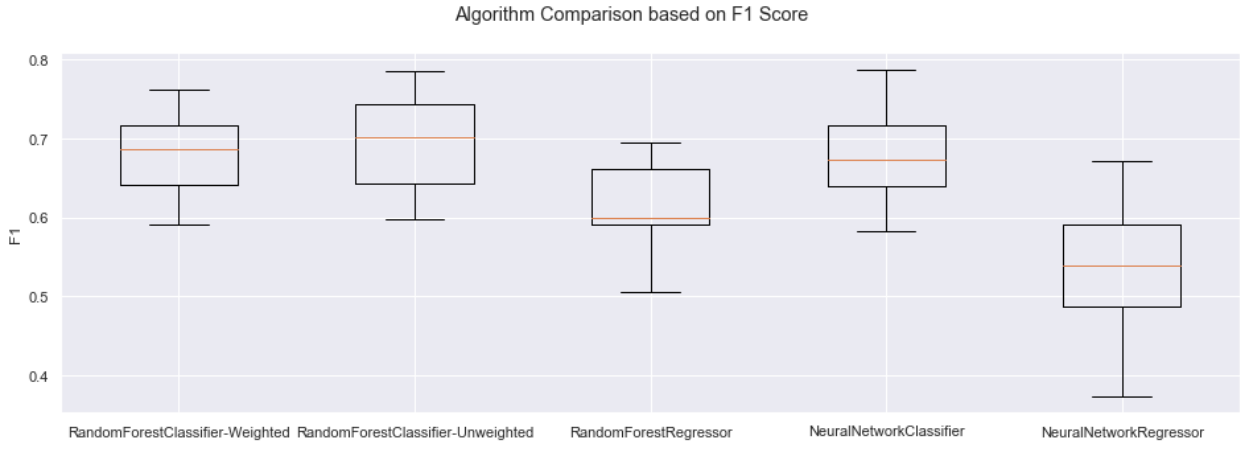


Figure 2.2. Comparison of AUC values from nested cross-validation of the Discriminative model.

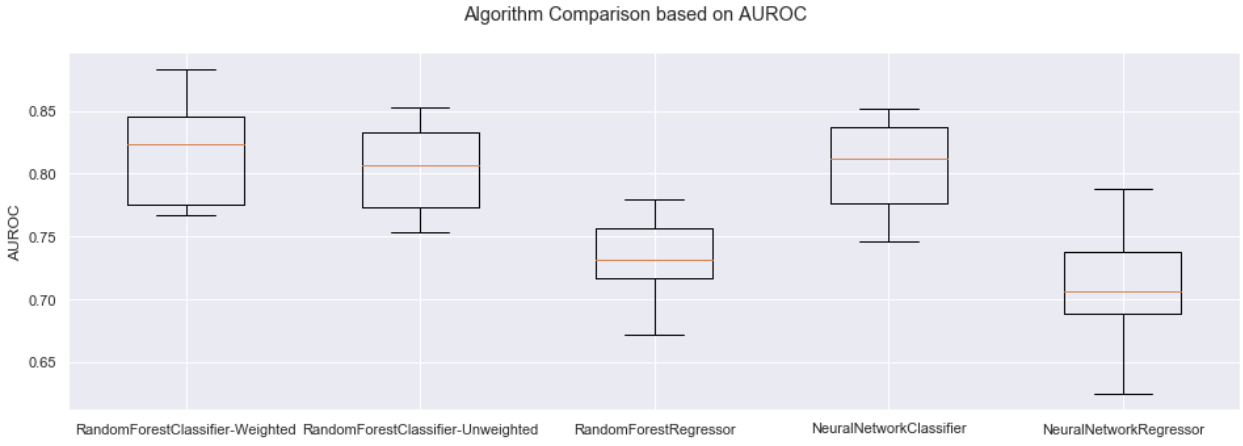
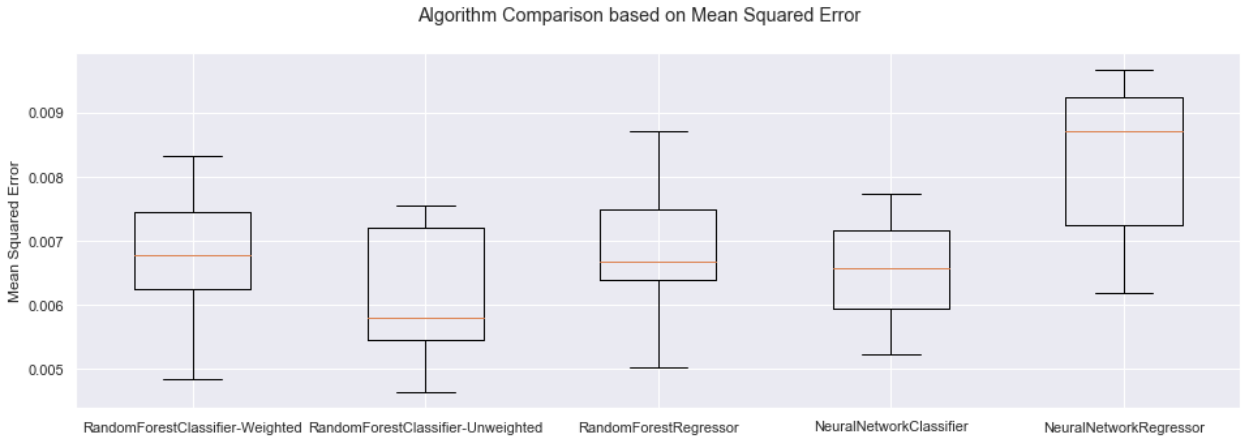


Figure 2.3. Comparison of mean squared error values from nested cross-validation of the Discriminative model.



Based on F1-scores, AUC, and mean squared error, the classification algorithms outperformed the regression algorithms, and the random forest classifiers (weighted and unweighted) outperformed the multilayer perceptron classifier. Given some overlapping performance (dependent on hyper-parameter choices), we trained each of the 5 models using the hyper-parameters that most frequently had the highest performance on the outer folds to serve as our best candidate models and evaluated their performance in the Validation Set. The weighted random forest classifier performed best and was designated the final Discriminative model.

We made two choices during the final Discriminative model development that potentially added information about the true labels. Therefore, we conducted a *post-hoc* sensitivity analysis. Specifically, in the final model, we specified the outcome label from the Generative model's predicted probability for the 51,712 records in the Training Set; however, for the additional 90 records in the Development Set, we specified the outcome label based on the manually-adjudicated determination. Additionally, we weighted samples during model fitting based on the absolute value of the Generative model probability's distance from 0.5. To achieve this, we created a vector containing a value ranging 0-0.5 to represent the Generative model's certainty of a record being a case versus control. Values closer to 0 represented low certainty (i.e., a random guess) while values closer to 0.5 represented greater certainty. We passed this vector as an argument to the scikit-learn implementation of the random forest algorithm, which was used to determine the penalty of mis-classifications (i.e., mis-classified predictions where the probabilistic outcome label was closer to 0.5 were penalized less than those closer to 0 or 1). We examined the influence of specifying the outcome label entirely from the Generative model (versus including the manually-adjudicated labels from the smaller Development Set) and weighting (versus not weighting) samples during the model fit.

In the last estimate of performance, we compared our final Discriminative model with the hold-out Test Set that was manually adjudicated via crowdsourcing. We used the Vanderbilt University Medical Center's Crowdsourcing Core services as an external review of the Test Set. The Crowdsourcing Core has an established workflow for assisting investigators in the describing desired outcomes for clinical chart reviews, recruiting and compensating qualified reviewers (known as "workers"), managing and displaying complex clinical data for review, and ensuring sufficient numbers of reviews to make a determination.[87] Workers completed the review in a two tasks, which were completely independent of the investigative team's activities. In the first task, workers evaluated whether the visit included an elective surgery. Visits without an elective surgery (i.e., those with no surgery or an emergent surgery) were excluded from

further review. In the second task containing only those visits with a confirmed elective surgery, workers evaluated whether respiratory depression occurred and whether it was likely due to opioid administration.

We received IRB approval for all activities involving human subjects.

Results

Labeling Functions in Training and Development Sets

After 5 rounds of LF creation and modification using our Training and Development data sets, we finalized our Generative model with 14 LFs (see Table 3 for final rules).

Table 2.3. Final LFs for identifying OIRD in the Generative model.

	Yes	**No**
Received naloxone (Narcan)?	CASE, if nearby keywords suggested naloxone administration was effective in reversing OIRD *or* CONTROL, if nearby keywords suggested naloxone administration was ineffective in reversing OIRD	CONTROL
The count of keywords suggesting naloxone *ineffectiveness* was greater than the count of keywords suggesting naloxone *effectiveness*?	CONTROL	CASE, if count > 0 *or* ABSTAIN, if no keywords present
Had an extended period (>= 4 days) of mechanical ventilation?	CONTROL	ABSTAIN
Had diagnostic codes for respiratory failure?	CONTROL, if mechanical ventilation also present	ABSTAIN
Absence of clinical notes with a title of “Respiratory Care”?	CONTROL	ABSTAIN
Keywords related to *narcotic overdose* were present?	CASE	ABSTAIN
Keywords related to *hypoxia* were present in clinical notes near variations of the word *opioid* or *narcotic*?	CASE	ABSTAIN
Keywords related to *decreasing opioids* were present?	CASE	ABSTAIN
Keywords related to *holding opioids* were present?	CASE	ABSTAIN
Keywords related *no pain meds* were present?	CONTROL	ABSTAIN
Keywords related to *altered mental status* were present?	ABSTAIN, if a confounding diagnosis (e.g., sepsis, myocardial infarction) present *or* CASE, if confounding diagnoses absent	ABSTAIN
Keywords related to *pinpoint pupils* were present?	CASE	ABSTAIN
The phrase “no acute events” was present?	ABSTAIN, if acute event keywords (e.g., “rapid response”, “altered mental status”) present *or* CONTROL, if acute event keywords absent	ABSTAIN
There were no keywords to support OIRD (e.g., hypoxia, rapid response, pinpoint pupils) present?	CONTROL	ABSTAIN

Validation Set Performance

In the Validation Set, the empirical accuracy of individual LFs ranged 0.47-1.00, the final Generative model achieved an accuracy of 0.83, an F1 score of 0.73, and an AUC of 0.96 (see Figure 4), and the final Discriminative model achieved an accuracy of 0.88, an F1 score of 0.80, and an AUC of 0.92 (see Figure 5). Performance of the final Discriminative model in the Validation Set was consistent with expected performance from the nested cross-validation process.

In the *post-hoc* sensitivity analysis, the Discriminative model trained with the removal of manually-adjudicated outcome labels from the Development Set (i.e., all outcome labels were produced by the Generative model's probabilistic labels) yielded the same accuracy, F1 score, and AUC values in the Validation Set. Conversely, the Discriminative model trained without sample weighting during the model fit yielded decreased accuracy (0.87), F1 score (0.79), and AUC (0.91) values in the Validation Set. During a review of record-level performance in the Validation Set, records with large a discrepancy between the predicted probabilities of Generative and Discriminative models primarily occurred when the Generative model indicated a probability close to 1 yet the manually-adjudicated label was "control." In sum, sample weighting during model fit improved overall model performance while the presence of manually-adjudicated labels corrected some records mis-classified as being a "case" in the Validation Set data.

Figure 2.4. Comparison of OIRD predicted probabilities from the Generative model with manually-adjudicated labels in Validation Set.

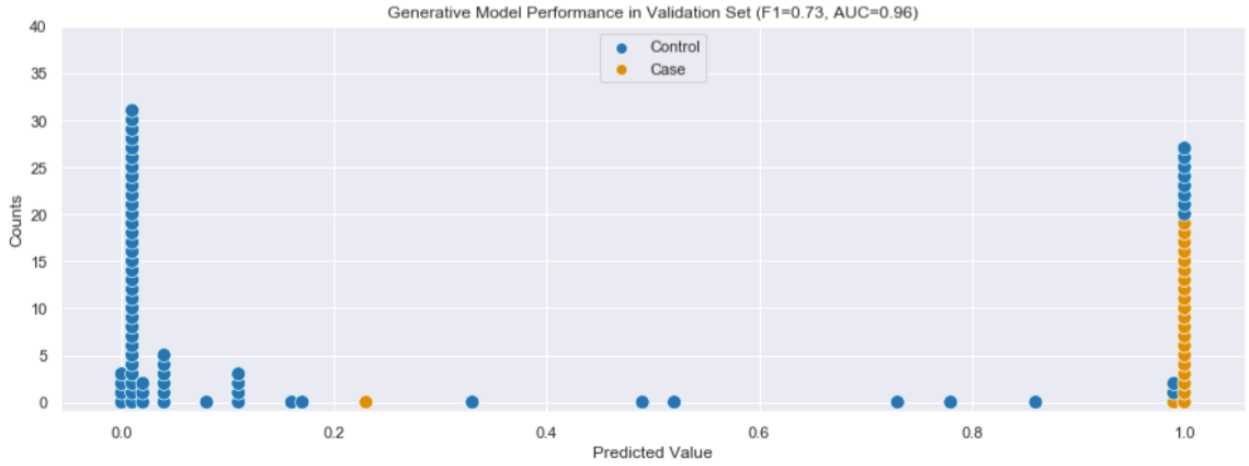
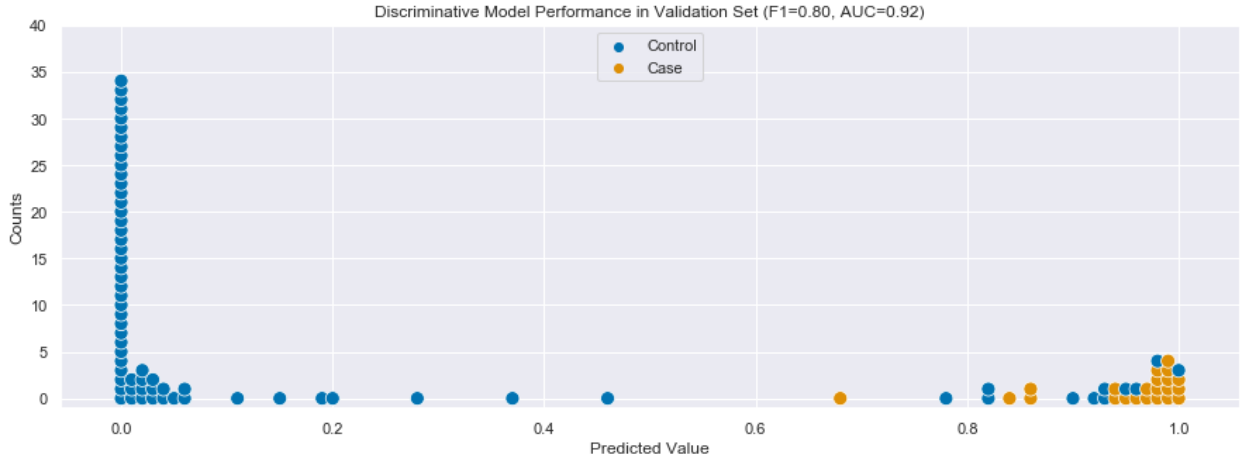


Figure 2.5. Comparison of OIRD predicted probabilities from the Discriminative model with manually-adjudicated labels in Validation Set.



Test Set Performance

In the first task, workers excluded 165 visits (21.6%) where the surgery was emergent, rather than elective. In the remaining 599 visits for the second task, workers determined OIRD was present in 5 (0.83%) visits. In the manually-adjudicated Test Set, the final Generative *and* Discriminative models achieved an accuracy of 0.977, an F1 score of 0.417, and an AUC of 0.988. Figures 7 and 8 illustrate the recall-precision curves for both models.

Figure 2.7. Recall-precision curve for Generative model in the Test Set.

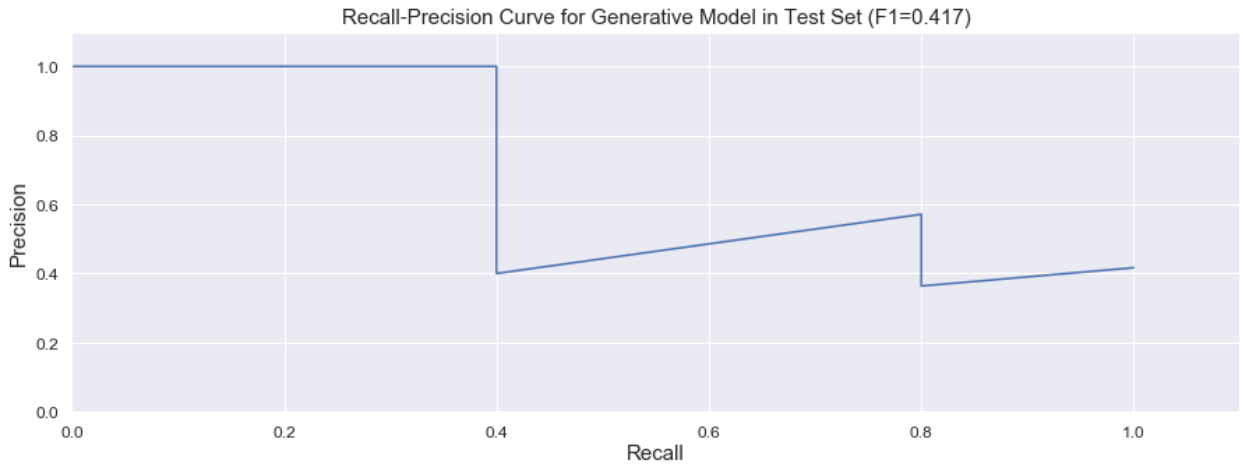
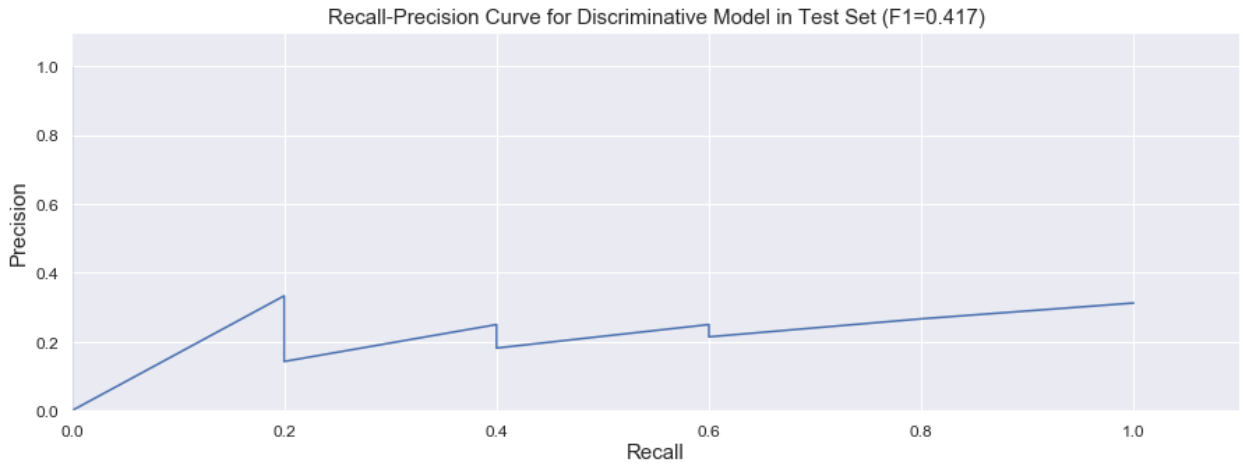


Figure 2.8. Recall-precision curve for the Discriminative model in the Test Set.



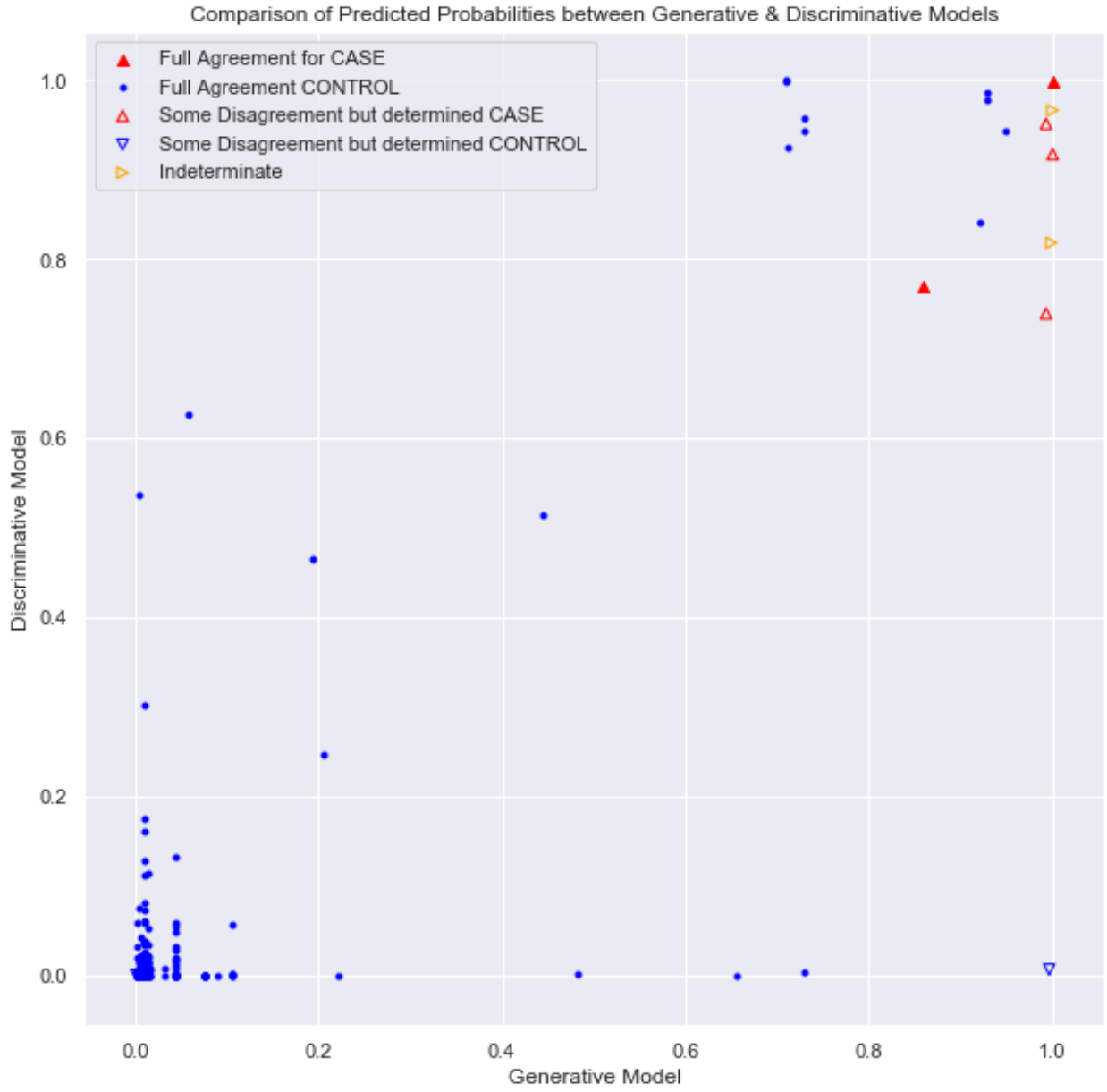
The Discriminative models used in the *post-hoc* sensitivity analysis for the Validation Set were associated with improved positive predictive values and F1 scores in the Test Set (see Table 4). The original AHRQ PSI-11 criteria performance in Test Set was lower than the Generative and Discriminative models (see Table 4) with 4 of the original 196 “cases” determined to be cases and 1 of the original 402 “controls” determined to be a case.

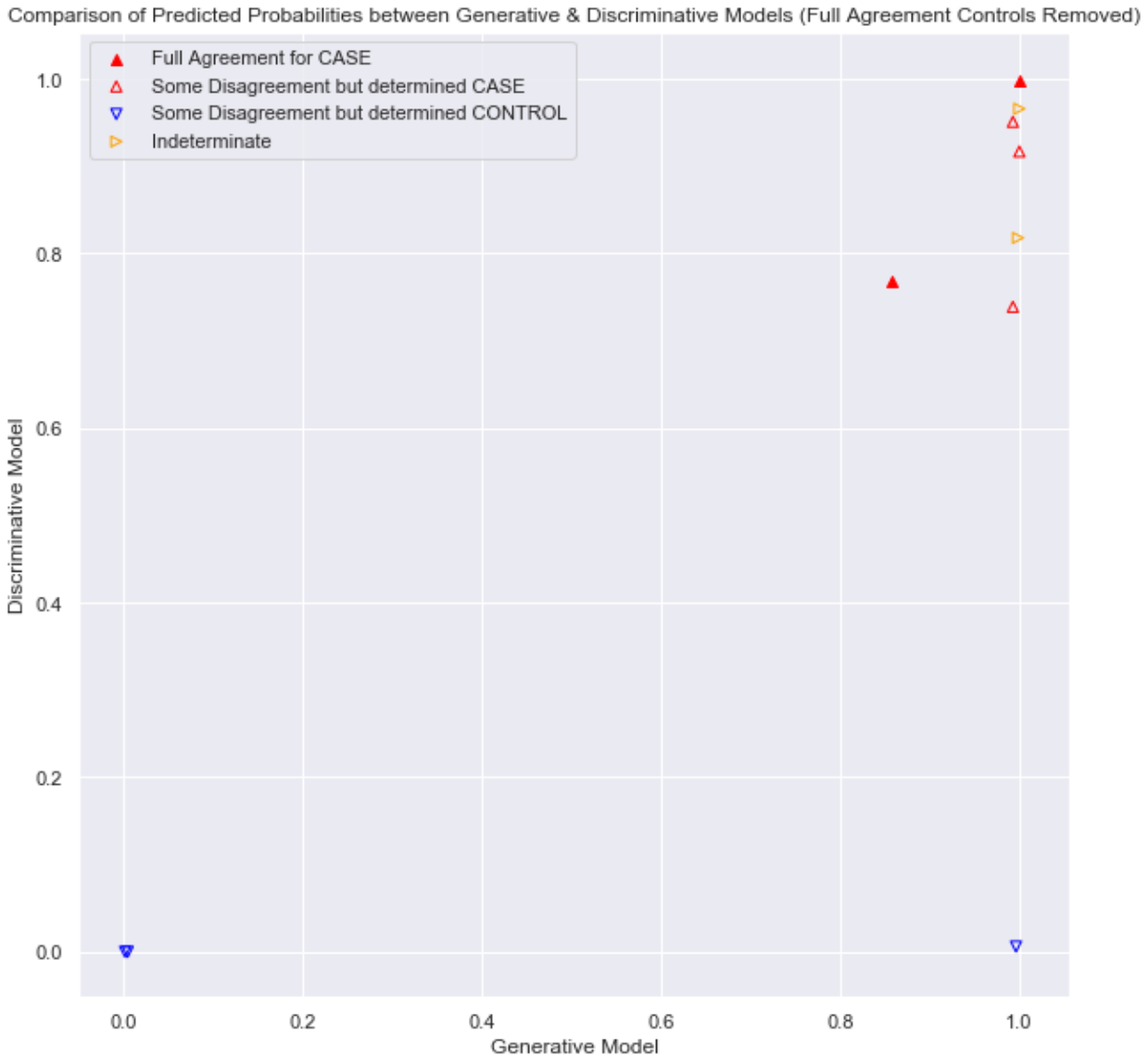
Table 2.4. Performance of all phenotyping approaches in Test Set.

	Sensitivity	**Specificity**	**Positive Predictive Value**	**Accuracy**	**AUC**	**F1 Score**
Generative Model	1.0	0.98	0.263	0.977	0.988	0.417
Final Discriminative Model (with weighting and some manual labels)	1.0	0.98	0.263	0.977	0.988	0.417
Discriminative Model without weighting	1.0	0.98	0.278	0.978	0.989	0.435
Discriminative Model without any manual labels	1.0	0.98	0.278	0.978	0.989	0.435
Discriminative Model without weighting or any manual labels	1.0	0.98	0.278	0.978	0.989	0.435
AHRQ PSI-11 Criteria	0.8	0.68	0.020	0.677	0.738	0.040

When examining the final Test Set status in the context of both the Generative and Discriminative models, all of those identified as a Case have a Generative model probability > 0.8 and a Discriminative model probability > 0.7 (see Figure 9). If one used these higher, joint thresholds, a revised scoring system would have an F1 score of 0.625.

Figure 2.9. Comparison of predicted probabilities between Generative and Discriminative models with final case/control status denoted. TOP: All results. BOTTOM: Visits determined to be a Control with full agreement on manual review are removed.





Review of Misclassified Patients

In the Validation Set, 11 of the 90 patients were classified as Cases based on the Discriminative model when the manual review (blinded to the Discriminative model's assigned probability) classified the patients as Controls. None of the patients were misclassified as controls. Table 4 contains the predicted probabilities along with comments from manual review of the Validation Set.

Table 2.4. Predicted probabilities and manual review comments from misclassified visits in the Validation Set.

Predicted Probability	Comments from Manual Review
1.0	Urethral cancer removal. Complex pain management for chronic cancer pain - always in pain but also became somnolent - no naloxone administration but suggested altered mental status.
0.98	Colostomy placement. Given naloxone & intubated after altered mental status - seems to be more like aspiration pneumonia. Naloxone only mentioned once & that patient became very anxious after administration.
0.96	Knee replacement. Rapid response team for respiratory compromise - likely due to metabolic acidosis or other causes. Naloxone was administered according the medication administration record but not in clinical notes (in fact one note suggests she had very little opioids).
0.95	No surgery described in clinical notes. Transferred from outside hospital for complex septicemia. Multiple notes discussed how the patient was given naloxone pre-hospital following opioid use at home.
0.93	Colectomy performed. Was somnolent & bradypneic requiring rapid response team - no effect from naloxone administration - likely due to alcohol withdrawal.
0.93	Dialysis patient presented to Emergency Department after blood cultures positive during dialysis & their operation was elective surgery for severely infected teeth. 1-month stay in the hospital. Lots of discussion regarding suboxone, high opioid use, & holding opioids. No evidence of OIRD during visit. Interestingly, the patient returned within 5 days of discharge with OIRD in the community.
0.92	Artificial hip irrigation & debridement. Coded & died after a complication with septic shock - no evidence of OIRD. Had been on oral naloxone.
0.90	No surgery described in clinical notes. Transferred from outside hospital for sepsis. Pulmonary note identified decreased respiratory drive on mechanical ventilation due to "delayed clearance of sedating meds" because they had ruled out other neurological etiologies of altered mental status. Naloxone had been administered, but this was via oral route and likely for constipation. Etiology of respiratory depression is unclear.
0.82	No surgery described in clinical notes. Transfer from outside hospital for sickle cell-related stroke. Had been taking high doses of narcotics at home.
0.82	Kidney & heart transplant. Didn't do well with extubation on post-operative day 1 & naloxone administration didn't help. Unlikely OIRD.
0.78	Pituitary tumor resection. No complications.

During a post-hoc manual review of the Test Set visits with high (>= 0.5) Discriminative model probabilities but labeled as Controls (n=14), the investigative team agreed with all crowdsourcing results and did not re-classify any Controls as Cases. However, one visit was deemed ambiguous/unclear by the crowdsourcing workers with one worker labeling the visit as a Case and one worker labeling the visit as a Control with no tie-breaker available. The

investigative team re-classified the visit from Unknown to Case. Table 5 contains the predicted probabilities along with comments from investigative team's post-hoc manual review of the Test Set visits with high Discriminative model probabilities among Control visits.

Table 2.5. Predicted probabilities and manual review comments from misclassified visits in the Test Set.

Predicted Probability	**Comments from Manual Review**
1.0	Hip replacement. No complications.
1.0	LVAD implant. Lengthy hospital stay with a discharge summary noting their "respiratory status remained tenuous".
0.99	Cystectomy for prostate cancer. Originally on room air, then increasing oxygen requirements and re-intubated on post-operative day 2 for unclear etiology, but not likely opioids.
0.98	Heart transplant. Very lengthy hospital stay and was intubated for a while.
0.97	Parathyroidectomy and thymectomy. Altered mental status that resulted in imaging evaluation where they received morphine and mental status worsened. Clinical notes reported some improvement with Narcan; however, OIRD seems unlikely given that they were tachypneic during that event.
0.96	Liver transplant. Improved gradually and uneventfully.
0.94	Liver transplant. Improved gradually and uneventfully.
0.94	Ileostomy takedown. Altered mental status of unknown origin. Seizure activity was originally assumed but no diagnostic evidence. Their morphine patient-controlled analgesia was making them sleepy, so it was discontinued. They received a couple doses of naloxone but no immediate improvement.
0.93	Partial nephrectomy for mass. Uneventful hospital course.
0.84	Fine needle aspiration and craniotomy for volumetric stereotaxy. Uneventful hospital course.
0.82	Pancreatojejunostomy for pancreatitis and hepatitis. Altered mental status and acute kidney injury that resulted in discontinuation of patient-controlled analgesia and naloxone administration. It appears sepsis was the complicating etiology rather than OIRD.
0.63	Percutaneous nephrolithotomy. Altered mental status with hypoxia and hypotension. Naloxone administered twice without improvement, and they ultimately died in the hospital.
0.54	Choledoduodenostomy. Uneventful hospital course.
0.51	Esophageal hernia repair. There were multiple mentions of naloxone in the medication lists from copy and paste of progress notes. They had post-operative complications involving being reintubated for hernia return and went to the Surgical ICU.

Discussion and Conclusion

We applied a data programming paradigm with the use of weak learners and heterogenous data types to the problem of identifying OIRD among post-operative adult patients. While our Generative model performed well in a small Validation Set, our

Discriminative model had lower performance in the larger, hold-out Test Set. Post-hoc review of misclassified visits from the Test Set provide insights into additional LFs that could be written to improve performance in future work. Notably, all of the confirmed Cases were identified by the Generative and Discriminative models. For rare outcomes, this finding is encouraging because it reduces the number of manual reviews needed by excluding visits/patients with low probabilities. As new patient records are added to our de-identified EHR database, we could score each record with the Discriminative model quickly and follow up with a manual review for records with high scores. While it would be possible to use the Generative model for scoring, it would be more challenging to incorporate data from external organizations (and similarly, to share the Generative model with external collaborators) due to the additional pre-processing steps required for applying LFs and creating a label matrix.

In our post-hoc sensitivity analysis of potential information added to the Discriminative model in the Validation Set, our results suggested sample weighting (based on the degree of uncertainty in the Generative model) improved overall performance and incorporating the outcome labels from manual adjudication corrected some mis-classification. This latter finding is likely due to the iterative enrichment of our Development Set and Validation Set with the top 20 Generative model probabilities as we developed LFs. Enriching both Sets with relatively homogenous records (i.e., the highest probabilities) and then building a Discriminative model with the combined Training and Development Sets resulted in added information that improved predictions in the Validation set. We did not see this added information influence performance in the hold-out Test Set where the Generative and Discriminative models performed similarly. However, we did observe improved performance in the Test Set (with respect to both positive predictive value and F1 score) of the unweighted model as well as removal of the manually-adjudicated labels. This observation suggests our final Discriminative model was slightly over-fit with a higher number of false positives.

Other biomedical studies have started to use the paradigm proposed by Snorkel (e.g.,

post-market medical device surveillance[88], extraction of pain levels from EHR notes[33]). Others' work using Snorkel suggests the Discriminative models perform better than Generative models,[33] so we hypothesized model performance on the hold-out Test Set would be high. What we found was that the two models performed differently, and there could be merit in considering both for creating outcome labels.

We initially followed the Snorkel developers' guidance for all steps in the labeling process but ultimately made some modifications, which we believe add to the literature. We abandoned the suggested context hierarchy[33] in favor of treating an entire visit as a single record/exemplar, which resulted in individual LF performance improvement. We also proposed a new method for Generative model hyper-parameter tuning by emphasizing the learned weights of the LFs rather than focusing on empirical accuracy, a modification which makes theoretical sense but should be examined more robustly in future studies.

Our work also has its limitations. Our data source did not identify the elective nature of its surgeries, so several non-elective surgeries were present. We attempted to overcome this limitation with the removal of visits where the surgical date occurred on the same day as an Emergency Department visit. Given some data sources will have direct access to this information, it is worth noting the limitation of data generated from a single organization in this study. Further, in the external manual review of the hold-out Test Set, the workers' first task was to remove non-elective surgeries. Another limitation of our work is a relative reduction in the potential data types included in LFs. For example, when exploring the effectiveness of naloxone administration, we attempted to incorporate the cosine similarity of vector embeddings of text data compared to examples of text suggesting naloxone effectiveness without success. Future studies could examine whether this contemporary natural language processing method improves LF performance. Similarly, clinical notes authored by nurses were not typically available in our data source. Although it is unlikely a nurse would document evidence of OIRD when a prescribing provider does not, that scenario could occur and should be examined in

future work. Finally, our iterative LF development process depended on enriching the Development Set and Validation Set based on the highest probabilities of candidate Generative models. We did not enrich our data sets for Control status (i.e., lower probabilities), but Control enrichment could easily be included depending on the clinical outcome under investigation.

In conclusion, we believe that a number of weak learners, when combined within a Snorkel framework, can facilitate identification of a complex outcome with a reduced number of manual chart reviews.

Chapter 3: Risk Prediction Modeling for Opioid-Induced Respiratory Depression

Introduction

Given the prior work in providing labels to a large observational data set, we can now use the labels to develop a clinical risk prediction model for opioid-induced respiratory depression (OIRD).

Methods

Design

We conducted a retrospective cohort analysis of post-operative patients to develop a predictive model for OIRD.

Sample & Setting

We collected data from post-operative adult patients in the de-identified electronic health record at Vanderbilt University Medical Center through process described in Chapter 2. However, rather than separate Training, Development, and Validation Sets, we combined these three sets into a single Training Set. The Test Set remained the same as described in Chapter 2.

Outcome

To define the OIRD outcome, we used our prior work from Chapter 2 where we developed a Discriminative Model to apply outcome labels to all records in the Training Set. Of the 51,812 visits in the Training Set, the Discriminative Model classified 594 (1.15%) visits as having OIRD. We used the manually-reviewed (i.e., crowdsourced) determinations as outcome labels for records in the Test Set. Of the 599 visits in the Test Set following removal of non-elective surgeries, 5 (0.83%) visits were classified as having OIRD.

Predictors

We sought to include a heterogenous set of predictor variables that represented the causal pathway of OIRD (see Discussion section). Due to data availability in our de-identified data source (see Limitations section), we had a limited feature set comprising: age on admission, serum creatinine level, billing diagnostic codes grouped into the top 15 categories from the Clinical Classification System (CCS),[89,90] American Society of Anesthesiologists' Physical Classification Systems class, and whether general anesthesia was administered.

We generated the CCS categories from versions 9 and 10 of the International Classification of Diseases (ICD). The CCS system, developed by the AHRQ as part of the Healthcare Cost and Utilization Project, maps individual ICD codes to fewer, clinically meaningful categories, which facilitates dimensionality reduction. ICD-9 codes can be mapped directly to CCS categories with software from AHRQ; however, the mapping of ICD-10 codes is still under development. To overcome this limitation, we leveraged the widely-used, hierarchical Systematized Nomenclature of Medicine (SNOMED) – Clinical Terminology[91] as an intermediary mapping vocabulary to first map ICD-10 codes to SNOMED codes and then determined whether an ICD-9 code could be mapped to the same SNOMED code. If no ICD-9 code mapped to the SNOMED code, we then looked for a match in the parent (i.e., one level up in the hierarchy) and grandparent (i.e., two levels up in the hierarch) SNOMED codes. If an ICD-9 code match was found in any of these three levels, we mapped the ICD-10 code to the associated CCS category. This process resulted in mapping 26,604 of 28,593 (93.0%) *unique* ICD codes and 4,278,308 of 4,398,328 (97.3%) *total* ICD codes found in our cohort to a CCS category. We reviewed a random subset of approximately 100 mappings, which revealed accurate mapping.

Data Pre-Processing

To account for missing data in the prediction features, we first imputed a value of "0" for administrative billing diagnostic codes – i.e., if missing, we assumed those patients did not have the associated diagnosis. For the remaining predictors, we applied the *IterativeImputer* function from scikit-learn[92] initialized with a median value and constraining minimum and maximum values to be bounded by the observed minimum and maximum values. To prepare for common machine learning algorithms, we scaled and centered the imputed features with the *StandardScaler* function from scikit-learn.[92]

Analysis

Following data pre-processing, we developed multiple machine learning algorithms available within scikit-learn to generate candidate prediction models. Our machine learning algorithms included: logistic regression (LR), linear discriminant analysis (LDA), k-nearest neighbors (KNN), classification and regression trees (CART), a random forest (RF), Gaussian Naïve Bayes (NB), and a multi-layer perceptron (NN).[12] All models attempted to predict a binary OIRD outcome at any point during hospitalization using predictor variable values available within the first eight hours of a hospital admission. We used 5-fold cross-validation to estimate performance without overfitting.[12] We evaluated model performance using both areas under the receiver operating characteristic curve and F1 scores.[17,19] Due to poor initial model performance with default hyper-parameters (see Results), we did not perform hyper-parameter tuning.

Results

Using default hyper-parameters from several machine learning algorithms within scikit-learn, we were unable to create a prediction model that performed better than chance. AUC scores ranged 0.50-0.62, and F1 scores ranged 0.00-0.04.

Figure 3.1. Area under the receiver operating characteristic curve values for several off-the-shelf machine learning algorithms.

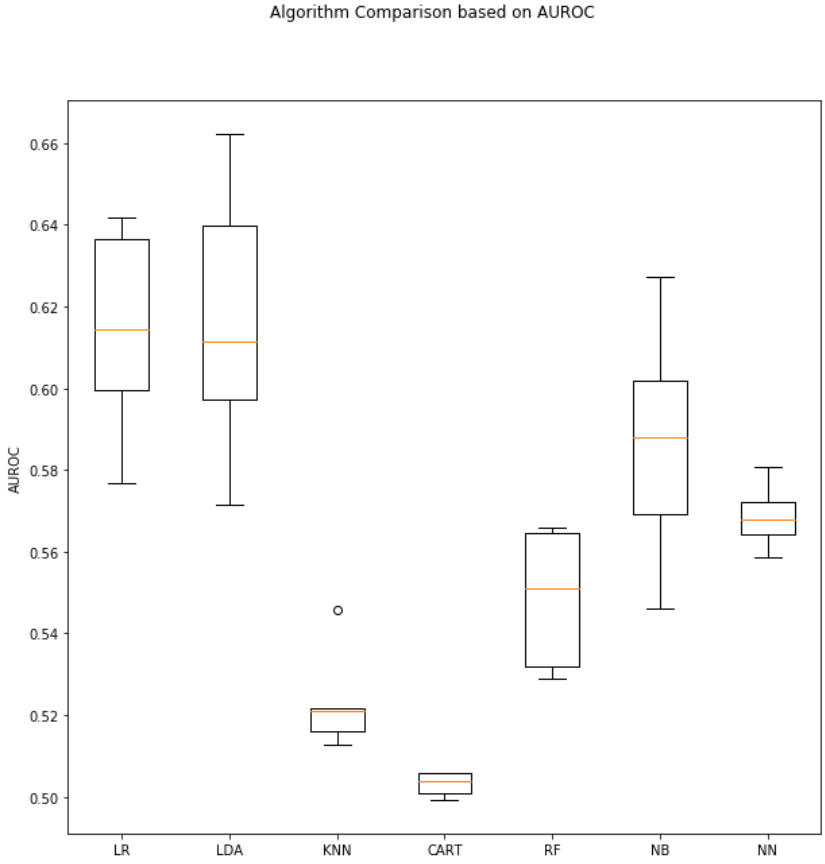
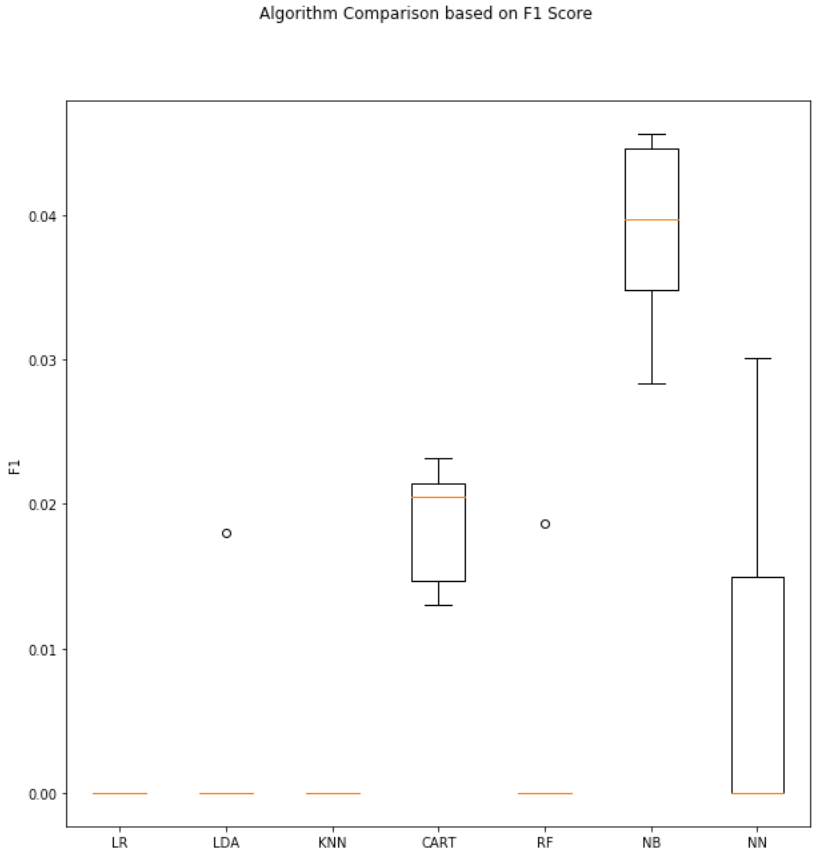


Figure 3.2. F1 score values for several off-the-shelf machine learning algorithms.



Discussion and Conclusion

We built an OIRD clinical prediction model with a small set of predictor variables. We were unable to develop a predictive model that performed better than chance; however, we now have an infrastructure for building a more robust model as new data are available.

Limitations

Based on a mechanistic understanding of OIRD, we originally intended to include the following additional predictor variables: temperature, heart rate respiratory rate, pulse oximetry, systolic blood pressure, diastolic blood pressure, a patient's pain level, amount of opioid received, information extracted from nurses' unstructured clinical note documentation, and

genetic data. We were unable to include the vital signs data due to a large amount of missingness (i.e., >50% of visits had missing vital signs in the pre-operative period). We were unable to include a patient's pain level because those structured data have not been reliably mapped from the operational/clinical data warehouse into the research data warehouses. Similarly, we were unable to include the amount of opioid received due to a data warehouse mapping error in which the mapping of drug exposures experienced a programming error for more than 18 months and had not been resolved at the time this work was completed. We were unable to include nurses' unstructured notes because they were not available in the data source (with the exception of Braden scoring system values). We were unable to include genetic data as predictors because there has not been a published GWAS for OIRD from which we could build a polygenic risk score, and the OIRD event rate in our data was not high enough to separately conduct a GWAS and then develop a polygenic risk score. All of these predictors should be considered for inclusion in future OIRD prediction studies.

Also aligning with our understanding the mechanistic underpinnings of OIRD, we had originally intended to develop the prediction model using a random forest within a discrete-time survival framework. This approach is in contrast to the method we used in the study where we used the last measured value occurring within the first eight hours of a hospital admission to predict OIRD at any point during hospitalization. In cases where variables can be measured at multiple time points (e.g., vital signs, laboratory values), a discrete-time survival framework permits the inclusion of repeated measures without performing case-control matching on an arbitrary time point. To prepare the data, each row of the matrix represents one timestamp for one patient. The matrix would be grouped by patient and sorted in chronological order; once a variable has a measured value, a last-one-carried-forward imputation would be applied. The outcome variable for each row would be a binary variable of whether the outcome occurred within the next 24 hours. This approach mimics clinical decision-making activities, mitigates the need for developing dependencies within patients, and has been a successful analysis

approach in others' work.[93] Further, developing a random forest model allows complex interactions that are more difficult to model in traditional regression. A random forest averages the results of many decision trees created by splitting a random selection of predictor variables in each tree.[12] Random forests are commonly used in the machine learning space and have demonstrated superior performance to other machine learning and traditional statistical approaches for outcomes related to inpatient clinical deterioration.[93] We plan to attempt this approach in the future when the data become available.

Chapter 4: Genome-Wide Association Study

Introduction

Given the prior work of developing an OIRD outcome label (i.e., phenotype) for a large observational data set and the lack of available genetic information for building OIRD predictive models, we sought to perform a genome-wide association study (GWAS) for OIRD.

Methods

Design

We performed a retrospective cohort analysis of post-operative patients to conduct a GWAS of OIRD.

Sample & Setting

As described in detail in Chapter 3, we collected data from de-identified EHR records at Vanderbilt University Medical Center. In addition to the EHR-based data, we used genetic information from BioVU. BioVU is a large DNA Databank with over 200,000 adult samples and almost 30,000 pediatric samples linked to detailed electronic health record data. Samples are obtained from leftover blood specimens collected during routine clinical care.[94] Samples were genotyped with Illumina MEGA-ex Array.

Phenotype Definition

To define the OIRD phenotype, we used our prior work from Chapter 2 where we developed a Discriminative Model to apply outcome labels to all records. We retained both the continuous probability value from the Discriminative Model as well as a binary (0/1) representation based on a threshold of 0.5. In records with a manually-adjudicated label (i.e.,

Development Set, Validation Set, and Test Set), we used the manually-adjudicated label (rather than the Discriminative Model label) as the binary phenotype.

Quality Control

To ensure adequate quality of the genetic data prior to a GWAS, we applied commonly-performed procedures based on recommendations by Marees et al.[95] and Reed et al.[96]

We removed single nucleotide polymorphisms (SNP) missing in > 5% of people and removed individuals missing more than 5% of SNPs. We assessed for sex discrepancies and, when discrepancies were present, we imputed sex based on SNP data. We originally observed 1,054 males and 1,134 females. The imputation procedure resulted in 1,055 males, 1,128 females, and 5 ambiguous individuals. Because none of the ambiguous/misclassified sexes were cases, we removed the ambiguous individuals from further analysis. We removed SNPs with a minor allele frequency <5% (in autosomal [1-22] SNPs only). We assessed Hardy-Weinberg equilibrium, but given the multi-ethnic cohort in our data, we did not exclude those 48,544 potentially problematic SNPs.

To identify and account for population substructure, we first extracted the variants present in our dataset from the 1000 genomes dataset, extracted the variants present in 1000 Genomes dataset from our dataset, and merged our data with the 1000 Genomes data set. We set the reference genome, resolved strand issues, and removed problematic SNPs from our data and the 1000 Genomes data. We performed multi-dimensional scaling (based on a principal component analysis) on our data anchored by the 1000 Genomes data and stored those features to serve as covariates in downstream regression models. We also used the scaling to estimate super-population association, and we excluded ethnic outliers that did not belong to the two largest sub-populations. This resulted in a European population comprising 1,927 individuals (with 14 cases) and an African population comprising 221 individuals (with 2 cases).

Finally, we removed individuals who deviated more than 3 standard deviations from the heterozygosity rate mean as well as individuals who were potentially related (first based on whether they were a case or control, then by keeping those with a higher call rate [i.e., a larger proportion of samples with a confident result from the genetic probe]).

Analysis

Following quality control procedures, we performed a GWAS with both a binary trait and a quantitative trait. We created the binary trait using a probability threshold of 0.5 from the Discriminative Model described in Chapter 2. We created the quantitative trait using the raw probability from the Discriminative Model. For both trait types, we calculated associations between all SNPs and the outcome with and without population sub-structure covariates. When including covariates, we used logistic regression for the binary trait and linear regression for the quantitative trait. To account for multiple testing, we performed unadjusted, Bonferroni-adjusted, and permutation-adjusted analyses. In the permutation-adjusted analyses, we attempted to include 1,000,000 permutations; due to computational feasibility, we restricted the number of permutations to 100,00 in the quantitative trait *without* covariates and 10,000 in the quantitative trait *with* covariates.

For post-hoc analyses, we created QQ plots to assess model quality and Manhattan plots to assist with interpretation of unadjusted and Bonferroni-adjusted results. Following the genome-wide association study, we compared the statistically significant SNPs to published studies and genome databases.

Results

Our genetic cohort comprised 17,271 patients who experienced a hospitalization (based on CPT codes beginning with 992). After restricting to AHRQ PSI-11 criteria (n=2,189) and

following quality control assessment, 14 cases and 1877 controls remained among those of European ancestry (the largest super-population in our cohort).

In the simple association studies with a binary phenotype, two single nucleotide polymorphisms (SNP) reached Bonferroni-adjusted statistical significance ($p < 0.05$), and one SNP neared significance (see Figure 1). With a continuous (i.e., probabilistic) phenotype, five SNPs reached significance, and one SNP neared significance (see Figure 3). In the regression models adjusted for population sub-structures as covariates, the binary phenotype was not associated with any statistically significant SNPs, but the continuous phenotype was associated with five significant SNPs and one near-significant SNP (see Figure 5). All Q-Q plots (see Figures 2, 4, and 6) suggest poorly-fit models. In the permutation-adjusted models, there were no statistically significant associations.

None of the significant SNPs had been previously identified from our literature review (see Chapter 1). An exploration of genes comprising (or closest to) the SNP revealed relatively frequent associations with previously published traits including anthropomorphic measures (e.g., height, waist circumference, body mass index), cognitive abilities, red blood cell measures, sex hormones, mental health disorders, Alzheimer's dementia, white blood cell counts or disorders (e.g., leukemias), lung function, diastolic blood pressure, scoliosis, and platelet counts. Table 1 contains details of each SNP finding, and Table 2 provides counts of the number of studies suggesting a trait is associated with a gene near a significant SNP in our study.

Figure 4.1. Manhattan plot of GWAS results with binary phenotype and unadjusted for covariates.

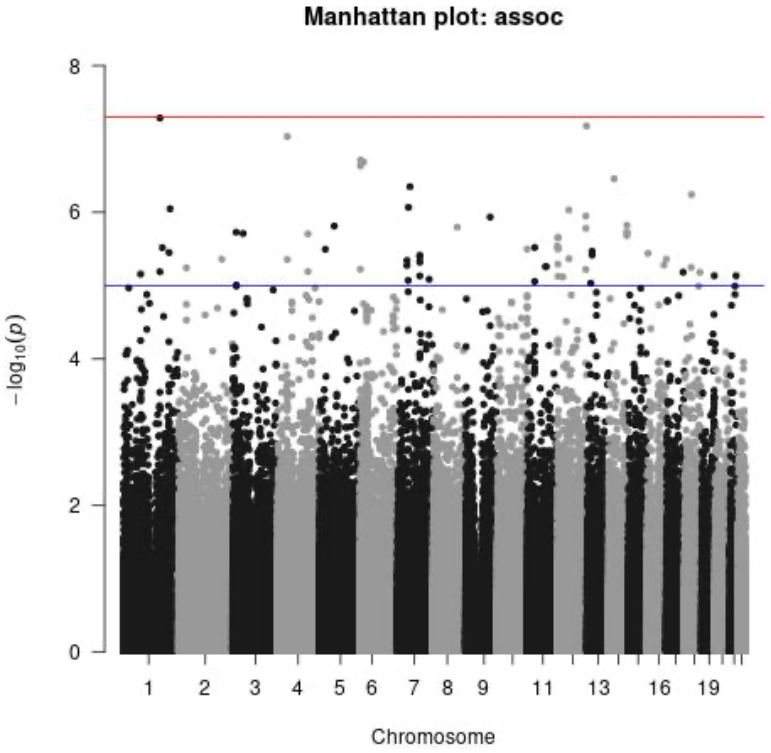


Figure 4.2. Q-Q plot of GWAS results with binary phenotype and unadjusted for covariates.

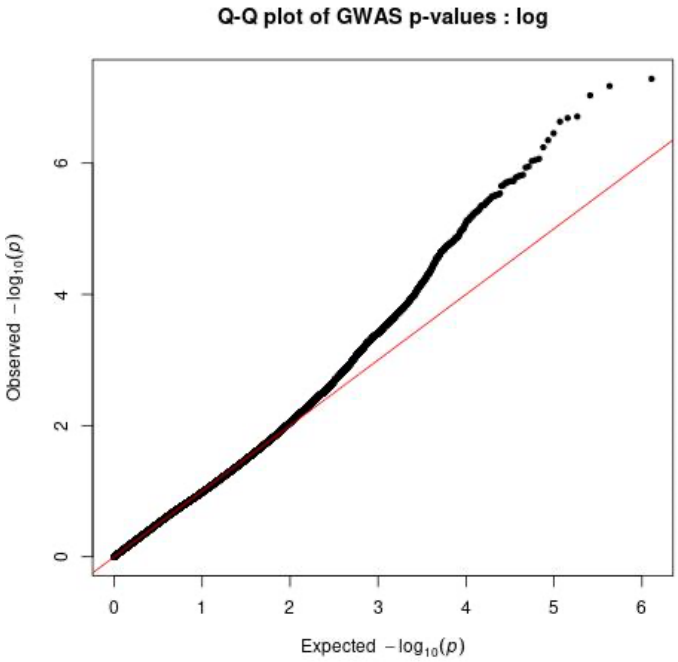


Figure 4.3. Manhattan plot of GWAS results with continuous phenotype and unadjusted for covariates.

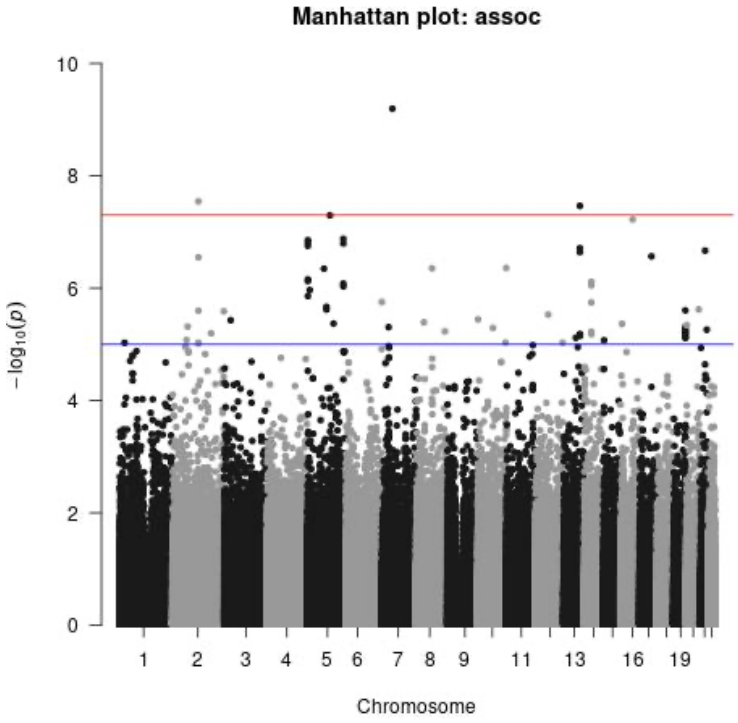


Figure 4.4. Q-Q plot of GWAS results with continuous phenotype and unadjusted for covariates.

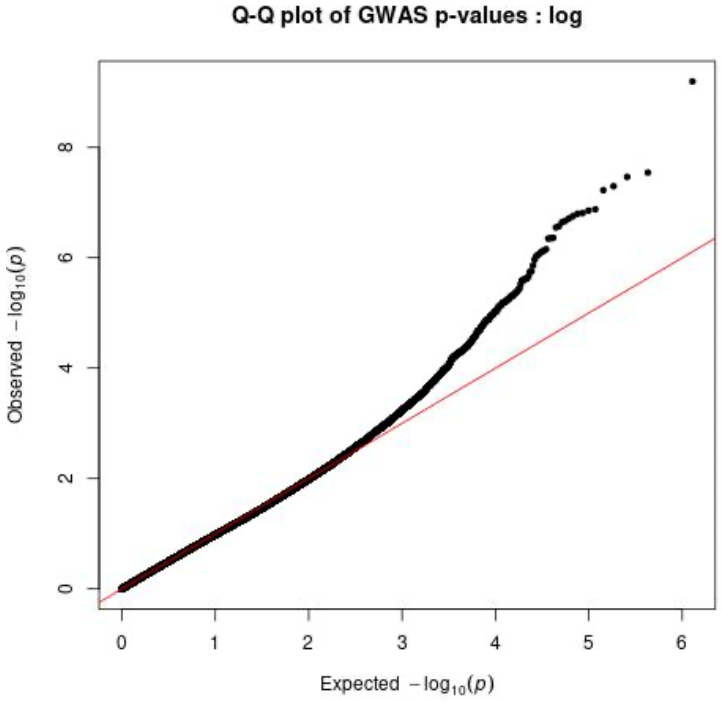


Figure 4.5. Manhattan plot of GWAS results with continuous phenotype sub-population covariate adjustment.

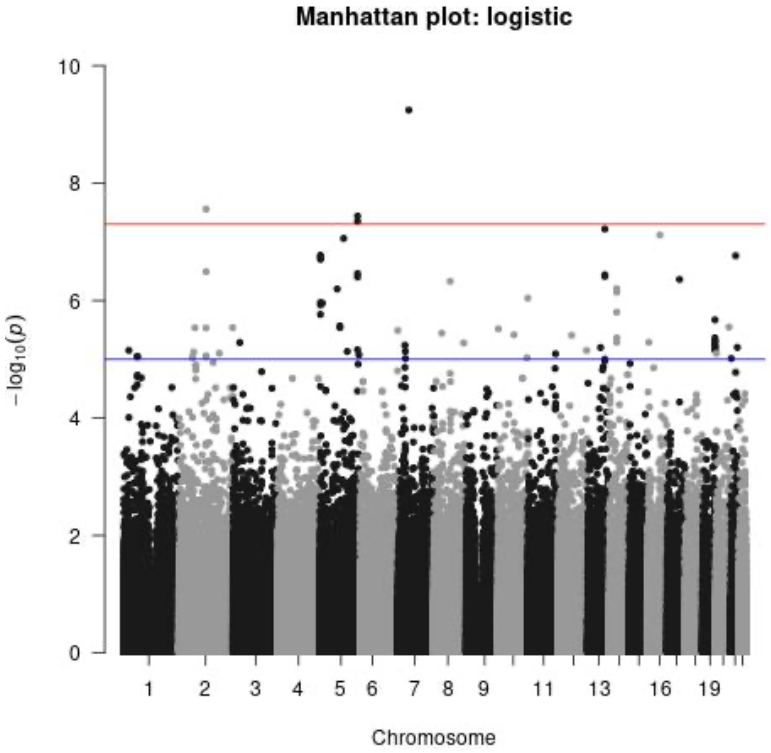


Figure 4.6. Q-Q plot of GWAS results with continuous phenotype with sub-population covariate adjustment.

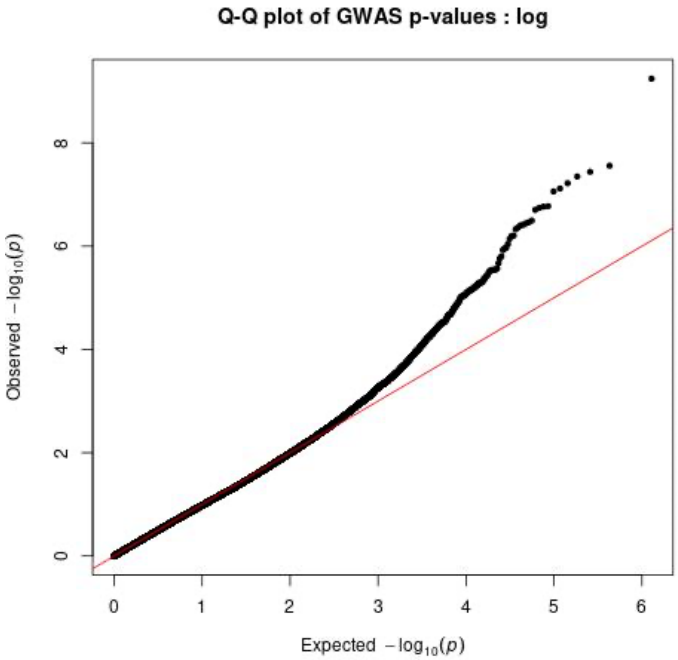


Table 4.1. GWAS results of SNP findings.

Phenotype	SNP	Illumina Probe	Covariate-Adjusted Regression?	p-value (Bonferroni-adjusted)	Chromosome	Closest Gene(s)	Within Gene?	Protein Encoding?
Binary	rs7416532	JHU_1.164445356	N	< 0.03336	1	PBX1	N	Y
						HMGB3P6	N	Processed Pseudogene
	rs73477358	JHU_12.131516369	N	< 0.043	12	ADGRD1	Y	Y
	rs511310	4:46240004-A-G	N	< 0.05967	4	GABRA2	N	Y
Quantitative	rs60757058	rs60757058	N	< 0.0004107	7	C7orf65	N	N
			Y	< 0.0003671		LINC01447	N	N
	rs7605011	rs7605011	N	< 0.01842	2	GLI2	Y	Y
			Y	< 0.0178				
	rs113513760	JHU_13.95615048	N	< 0.02209	13	LINC00557	N	N
			Y	< 0.04921				
	rs114457728	rs114457728	N	< 0.03251	5	EFNA5	Y	Y
			Y	< 0.0561				
	rs72778715	rs72778715	N	< 0.03841	16	CPNE2	Y	Y
			Y	< 0.04921				
	rs90213	exm2270266	N	< 0.0855	5	DOCK2	Y	Y
			Y	< 0.02343				
		rs90213	Y	< 0.0288				

Note: Closest gene(s) identified from https://www.ncbi.nlm.nih.gov/snp/

Table 4.2. Counts of the number of studies suggesting a trait is associated with a gene near a significant SNP in our study.

	Anthropomorphic Measures	Cognitive Ability	Red Blood Cells	Sex Hormones	Mental Health Disorders	Alzheimer's Dementia	White Blood Cell Counts or Disorders	Lung Function	Diastolic BP	Scoliosis	Platelets	Other Reported Traits (https://www.ebi.ac.uk/gwas/home)
Total	**35**	**23**	**15**	**11**	**9**	**7**	**6**	**5**	**4**	**3**	**3**	
Gene												
PBX1	13	1	3	6			4		1		3	ACEI cough (1), Hepatitis B (1), Night sleep (1), Intrinsic epigenetic age acceleration (1), Pre-treatment HIV viral load (1)
HMGB3P6								1		1		
ADGRD1	6			2	1	1				1		Systolic BP (1), Serum phosphate levels (3), Amygdala volume (1), Macrophage migration inhibitory factor levels (1), Longevity (1), ALS (1), RR interval [heart rate] (1), Gut microbiota relative abundance (1), Neurofibrillary tangle (1)
GABRA2		1				1		2				Risk tolerance & adventerousness (2), Epilepsy (2), Protein quantitative trait loci [liver] (1), Age at diagnosis of Type 1 diabetes (1), Mononucleosis (1)
C7orf65									1			Corneal curvature (3), Refractive error (1), ALT in excessive ETOH consumption (1)

LINC01447			12						1			
GLI2	13			3				1	1			eGFR (1), Acne (1)
LINC00557						5	1					Paternal language impairment (1)
EFNA5	2	17			7			1				GGT levels (6), ALT levels (3), Smoking status (5), Brain region volumes (4), Household income (1), Metabolite levels (1), Chronotype (1), Age at first sexual intercourse (1), Daytime nap (1), CTACK levels (1), Non-del(5q) myelodysplastic syndromes (1), Sedentary behavior duration (1), AST levels (1), Reaction time (1), Number of children ever born (1)
CPNE2												HDL levels (3), Apolipoprotein A1 levels (2), Oligosaccharide concentration of human milk (2), HDL interaction with short sleep time (1)
DOCK2	1	4			1		1			1		Placental abruption (2), PLT-derived growth factor BB levels (1), Pneumococcal bacteremia (1), Coronary artery aneurysm in Kawasaki disease (1), Protein

												quantitative trait loci [leptin] (1), IgG glycosylation (1), Age-related hearing impairment with SNP x SNP interaction (1)

Discussion and Conclusion

We conducted a GWAS for binary and continuous representations of OIRD. While the GWAS yielded potentially informative associations, the findings should be interpreted with caution due to the small sample size (particularly the number of cases) as evidenced by poor model fits on Q-Q plots and lack of significant associations in permutation-adjusted models. We are encouraged by the increased the number of statistically significant associations when using a continuous (rather than binary) phenotype, which theoretically facilitates less information loss during phenotype development.

In the future, we plan to gather additional samples from Vanderbilt's genetic biobank to increase our sample size. With a larger sample, we should have the ability to conduct a more robust GWAS followed by the development of a polygenic risk score that could be included in OIRD risk prediction models.

Chapter 5: Conclusion

In this thesis work, we applied a data programming paradigm with the software system Snorkel[33] to develop outcome labels for (i.e., opioid-induced respiratory depression [OIRD]) and attempted to use those labels to build a predictive model. The use of Snorkel to phenotype OIRD in a large observational data set was successful, particularly with its 100% sensitivity in a hold-out Test Set. This method opens new opportunities for identifying rare, incompletely ascertainable outcomes in large clinical data sets. Although the F1 score suggested only moderate overall performance, the high sensitivity of Snorkel's predictions combined with the low prevalence of OIRD results in significantly fewer manual chart reviews (compared to not using Snorkel) necessary to apply phenotypes to the entirety of a large data set.

We made two modifications to the data programming paradigm within Snorkel that should continue to be examined in other domains. We first collapsed the context hierarchy in order to treat a patient's longitudinal data during their hospital encounter as a single row of data, which was not only a simpler method for applying labeling functions but also resulted in greater empirical accuracy. We also introduced an approach to hyper-parameter tuning of the Generative model that relied on the learned weights of the labeling functions rather than the empirical accuracy of the Generative model in a validation set. In this work, we used a single reviewer to determine which rank ordering had the greatest face validity for clinical relevance. Additional work is needed to explore whether a more reliable and valid approach for determining the most appropriate ranking is possible. In the future, we plan to apply Snorkel to other clinical domains to evaluate performance and explore under what conditions (e.g., data types, data quality, number of labeling functions, scientific programming experience of research investigators) Snorkel performs well.

We were unsuccessful in building an OIRD prediction model due to a number of missing variables. This problem could be resolved in future work once we have access to the missing

data. Notably, no published OIRD clinical prediction models were available at the beginning of this work; however, during the execution of this thesis work, two relevant studies were published. One group used a medical device to continuously monitor the exhaled carbon dioxide as well as oxygenation status of patients for patients at high OIRD risk.[97] To identify high-risk patients who might benefit from continuous monitoring, they developed a predictive model considering the following *candidate* predictors: age, sex, body mass index, smoking status, acute bronchitis, aortic aneurysm, aortic valve disease, asthma, cerebral aneurysm, chronic bronchitis, heart failure, chronic obstructive pulmonary disease, coronary artery disease, diabetes mellitus, hypertension, kidney failure, liver failure, mitral valve disease, myocardial infarction, known or suspected sleep disorders, peripheral vascular disease, pulmonary hypertension, sepsis, stroke, transient ischemic attack, number of different opioids, opioid naivety, high risk surgery, open surgery, and duration of surgery. The *final* predictor list comprised only age, sex, opioid naivety, sleep disorders, and chronic heart failure. Using their predictive model, they applied their continuous monitoring device to high-risk patients, and they found the device to be accurate. A separate study of 60 patients also used continuous monitoring and developed a model to predict OIRD in the immediate post-operative recovery period, but its limited sample size and lack of methodological detail make it difficult to critically evaluate the findings.[98] We plan to use these predictors as candidate features for consideration in future prediction model work. Once we have the necessary data to build a more robust OIRD predictive model, we will compare our findings to these published studies.

Our genome-wide association study (GWAS) for OIRD identified a few statistically-significant associations. However, the permutation-adjusted analysis results suggested these associations were spurious relationships (likely due to the small sample size). We plan to gather additional samples in the future for conducting a GWAS with greater power. Notably, though, our quantitative representation of OIRD yielded more statistically-significant associations than the binary representation, which is logical given that a continuous distribution has greater

statistical power than a binary representation. While this could also be spurious, it would be worth examining the influence of quantitative (i.e., continuous level) measures of traditionally-binary traits in future studies to determine whether new genetic insights are uncovered.

From a personal perspective in my journey as a scientist, I gained greater familiarity with managing and analyzing large data sets. I learned how to apply noisy labels for the purpose of phenotyping while implementing and evaluating a novel framework that has additional applications for the biomedical informatics community. I learned more about the representation of standardized concepts using the Observational Medical Outcomes Partnership during the pre-processing and feature engineering phases of the study. I made an incremental advancement to using the common data model to map International Classification of Diseases (10^{th} version) to the Clinical Classification System categories based on existing maps within the 9^{th} version of the International Classification of Diseases and the Systematized Nomenclature of Medicine (SNOMED) – Clinical Terminology. These skills have already been leveraged in other aspects of my research activities and research grant applications. As a lifelong learner, I look forward to continuing my learning journey beyond this thesis work.

REFERENCES

1. Wilson PW, Castelli WP, Kannel WB. Coronary risk prediction in adults (the Framingham Heart Study). The American journal of cardiology. 1987;59(14):G91-G4.
2. Pencina MJ, D'Agostino Sr RB, Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the Framingham Heart Study. Circulation. 2009;119(24):3078-84.
3. Adler-Milstein J, Jha AK. HITECH Act drove large gains in hospital electronic health record adoption. Health affairs. 2017;36(8):1416-22.
4. U.S. Department of Health and Human Services. Health IT Adoption Rates [updated January 15, 2013. Available from: http://www.healthit.gov/policy-researchers-implementers/health-it-adoption-rates.
5. Centers for medicare and Medicaid Services. Promoting Interoperability Programs 2021 [Available from: https://www.cms.gov/regulations-and-guidance/legislation/ehrincentiveprograms.
6. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health affairs. 2014;33(7):1123-31.
7. Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing electronic health care predictive analytics: considerations and challenges. Health Aff (Millwood). 2014;33(7):1148-54.
8. Weil AR. Big data in health: a new era for research and patient care. Health affairs. 2014;33(7):1110.
9. Collins G, Le Manach Y. Multivariable risk prediction models: it's all about the performance. Anesthesiology. 2013;118(6):1252-3.
10. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC medical research methodology. 2014;14:40.
11. Harrell F. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ed. New York, NY: Springer; 2015.
12. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning : data mining, inference, and prediction. 2nd ed. New York, NY: Springer; 2009. xxii, 745 p. p.
13. Agarwal V, Podchiyska T, Banda JM, et al. Learning statistical models of phenotypes using noisy labeled training data. Journal of the American Medical Informatics Association : JAMIA. 2016;23(6):1166-73.
14. Szepesvári C. Algorithms for reinforcement learning. Synthesis lectures on artificial intelligence and machine learning. 2010;4(1):1-103.
15. Singh A, Thakur N, Sharma A, editors. A review of supervised machine learning algorithms. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom); 2016: Ieee.
16. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making. 2019;19(1):1-16.
17. Liu X, Anstey J, Li R, Sarabu C, Sono R, Butte AJ. Rethinking PICO in the Machine Learning Era: ML-PICO. Applied clinical informatics. 2021;12(2):407-16.
18. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York, NY: Springer; 2009.
19. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.
20. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. Journal of the American Medical Informatics Association : JAMIA. 2013;20(e1):e147-54.

21. Overby CL, Pathak J, Gottesman O, et al. A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury. Journal of the American Medical Informatics Association : JAMIA. 2013;20(e2):e243-52.
22. Alzoubi H, Alzubi R, Ramzan N, West D, Al-Hadhrami T, Alazab M. A Review of Automatic Phenotyping Approaches using Electronic Health Records. Electronics-Switz. 2019;8(11).
23. Bastarache L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. Annual Review of Biomedical Data Science. 2021;4:1-19.
24. Yu S, Ma Y, Gronsbell J, et al. Enabling phenotypic big data with PheNorm. Journal of the American Medical Informatics Association : JAMIA. 2018;25(1):54-60.
25. Liao KP, Sun J, Cai TA, et al. High-throughput multimodal automated phenotyping (MAP) with application to PheWAS. Journal of the American Medical Informatics Association : JAMIA. 2019;26(11):1255-62.
26. Zhang Y, Cai T, Yu S, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nat Protoc. 2019;14(12):3426-44.
27. Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH, Elhadad N. Learning probabilistic phenotypes from heterogeneous EHR data. Journal of biomedical informatics. 2015;58:156-65.
28. Sinnott JA, Cai F, Yu S, et al. PheProb: probabilistic phenotyping using diagnosis codes to improve power for genetic association studies. Journal of the American Medical Informatics Association : JAMIA. 2018;25(10):1359-65.
29. Aslam JA, Decatur SE. On the sample complexity of noise-tolerant learning. Inform Process Lett. 1996;57(4):189-95.
30. Simon HU. General bounds on the number of examples needed for learning probabilistic concepts. J Comput Syst Sci. 1996;52(2):239-54.
31. Halpern Y, Horng S, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. Journal of the American Medical Informatics Association : JAMIA. 2016;23(4):731-40.
32. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Jt Summits Transl Sci Proc. 2017;2017:48-57.
33. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Re C. Snorkel: rapid training data creation with weak supervision. VLDB J. 2020;29(2):709-30.
34. Yang X, Bian J, Hogan WR, Wu Y. Clinical concept extraction using transformers. Journal of the American Medical Informatics Association : JAMIA. 2020;27(12):1935-42.
35. Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. Journal of biomedical informatics. 2017;73:14-29.
36. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. Artificial intelligence in medicine. 2016;71:57-61.
37. de Vries EN, Ramrattan MA, Smorenburg SM, Gouma DJ, Boermeester MA. The incidence and nature of in-hospital adverse events: a systematic review. Quality & safety in health care. 2008;17(3):216-23.
38. Agency for Healthcare Research and Quality. Patient Safety Indicators v6.0 ICD-9-CM Benchmark Data Tables. Rockville, MD2017.
39. Covidien. Respiratory Compromise is Common, Costly and Deadly. Boulder, CO; 2014.
40. Nelson LS, Juurlink DN, Perrone J. Addressing the Opioid Epidemic. JAMA : the journal of the American Medical Association. 2015;314(14):1453-4.
41. Psaty BM, Merrill JO. Addressing the Opioid Epidemic - Opportunities in the Postmarketing Setting. N Engl J Med. 2017;376(16):1502-4.
42. Friedmann PD, Andrews CM, Humphreys K. How ACA Repeal Would Worsen the Opioid Epidemic. New Engl J Med. 2017;376(10).

43. Cashman JN, Dolin SJ. Respiratory and haemodynamic effects of acute postoperative pain management: evidence from published data. British journal of anaesthesia. 2004;93(2):212-23.
44. Gupta K, Prasad A, Nagappa M, Wong J, Abrahamyan L, Chung FF. Risk factors for opioid-induced respiratory depression and failure to rescue: a review. Curr Opin Anaesthesiol. 2018;31(1):110-9.
45. Chidambaran V, Olbrecht V, Hossain M, Sadhasivam S, Rose J, Meyer MJ. Risk predictors of opioid-induced critical respiratory events in children: naloxone use as a quality measure of opioid safety. Pain medicine (Malden, Mass). 2014;15(12):2139-49.
46. Agency for Healthcare Research and Quality. Patient Safety Indicator 11 (PSI 11) Postoperative Respiratory Failure Rate (ICD-9-CM Version 6.0) 2017 [Available from: https://www.qualityindicators.ahrq.gov/Downloads/Modules/PSI/V60-ICD09/TechSpecs/PSI_11_Postoperative_Respiratory_Failure_Rate.pdf.
47. Agency for Healthcare Research and Quality. Patient Safety Indicator 11 (PSI 11) Postoperative Respiratory Failure Rate (ICD-10-CM v2018) 2018 [Available from: https://www.qualityindicators.ahrq.gov/Downloads/Modules/PSI/V2018/TechSpecs/PSI_11_Postoperative_Respiratory_Failure_Rate.pdf.
48. Borzecki AM, Cevasco M, Chen Q, Shin M, Itani KM, Rosen AK. How valid is the AHRQ Patient Safety Indicator "postoperative physiologic and metabolic derangement"? Journal of the American College of Surgeons. 2011;212(6):968-76 e1-2.
49. Henry LR, Minarich MJ, Griffin R, et al. Physician derived versus administrative data in identifying surgical complications. Fact versus Fiction. Am J Surg. 2019;217(3):447-51.
50. Nguyen MC, Moffatt-Bruce SD, Strosberg DS, Puttmann KT, Pan YL, Eiferman DS. Agency for Healthcare Research and Quality (AHRQ) Patient Safety Indicator for Postoperative Respiratory Failure (PSI 11) does not identify accurately patients who received unsafe care. Surgery. 2016;160(4):858-68.
51. Romano PS, Mull HJ, Rivard PE, et al. Validity of selected AHRQ patient safety indicators based on VA National Surgical Quality Improvement Program data. Health services research. 2009;44(1):182-204.
52. Utter GH, Cuny J, Sama P, et al. Detection of postoperative respiratory failure: how predictive is the Agency for Healthcare Research and Quality's Patient Safety Indicator? Journal of the American College of Surgeons. 2010;211(3):347-54 e1-29.
53. Dana-Farber Harvard Cancer Center. Genotyping and Genetics for Population Sciences Core 2017 [Available from: http://www.dfhcc.harvard.edu/research/core-facilities/genotyping-and-genetics-for-population-sciences/pricing/.
54. Cincinnati Children's Hospital Medical Center. SNP Genotyping 2017 [Available from: https://dna.cchmc.org/www/snpgen_main.php.
55. Children's Hospital of Philadelphia. Center for Applied Genomics: Pricing and Services 2017 [Available from: https://caglab.org/index.php/for-researchers/pricing-and-services.html.
56. Centers for Disease Control and Prevention. National Hospital Discharge Survey 2012 [Available from: https://www.cdc.gov/nchs/nhds/nhds_tables.htm#number.
57. Dahan A, Sarton E, Teppema L, et al. Anesthetic potency and influence of morphine and sevoflurane on respiration in mu-opioid receptor knockout mice. Anesthesiology. 2001;94(5):824-32.
58. Romberg R, Sarton E, Teppema L, Matthes HW, Kieffer BL, Dahan A. Comparison of morphine-6-glucuronide and morphine on respiratory depressant and antinociceptive responses in wild type and mu-opioid receptor deficient mice. British journal of anaesthesia. 2003;91(6):862-70.
59. Zhang Y, Wang D, Johnson AD, Papp AC, Sadee W. Allelic expression imbalance of human mu opioid receptor (OPRM1) caused by variant A118G. J Biol Chem. 2005;280(38):32618-24.

60. Kroslak T, Laforge KS, Gianotti RJ, Ho A, Nielsen DA, Kreek MJ. The single nucleotide polymorphism A118G alters functional properties of the human mu opioid receptor. J Neurochem. 2007;103(1):77-87.
61. Beyer A, Koch T, Schroder H, Schulz S, Hollt V. Effect of the A118G polymorphism on binding affinity, potency and agonist-mediated endocytosis, desensitization, and resensitization of the human mu-opioid receptor. J Neurochem. 2004;89(3):553-60.
62. Mura E, Govoni S, Racchi M, et al. Consequences of the 118A>G polymorphism in the OPRM1 gene: translation from bench to bedside? J Pain Res. 2013;6:331-53.
63. Walter C, Lotsch J. Meta-analysis of the relevance of the OPRM1 118A>G genetic variant for pain treatment. Pain. 2009;146(3):270-5.
64. Madadi P, Hildebrandt D, Gong IY, et al. Fatal hydrocodone overdose in a child: pharmacogenetics and drug interactions. Pediatrics. 2010;126(4):e986-9.
65. Madadi P, Sistonen J, Silverman G, et al. Life-threatening adverse events following therapeutic opioid administration in adults: is pharmacogenetic analysis useful? Pain research & management : the journal of the Canadian Pain Society = journal de la societe canadienne pour le traitement de la douleur. 2013;18(3):133-6.
66. Stamer UM, Stuber F, Muders T, Musshoff F. Respiratory depression with tramadol in a patient with renal impairment and CYP2D6 gene duplication. Anesthesia and analgesia. 2008;107(3):926-9.
67. Chidambaran V, Sadhasivam S, Mahmoud M. Codeine and opioid metabolism: implications and alternatives for pediatric pain management. Curr Opin Anaesthesiol. 2017;30(3):349-56.
68. Owusu Obeng A, Hamadeh I, Smith M. Review of Opioid Pharmacogenetics and Considerations for Pain Management. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy. 2017.
69. Niesters M, Overdyk F, Smith T, Aarts L, Dahan A. Opioid-induced respiratory depression in paediatrics: a review of case reports. British journal of anaesthesia. 2013;110(2):175-82.
70. Overdyk F, Dahan A, Roozekrans M, van der Schrier R, Aarts L, Niesters M. Opioid-induced respiratory depression in the acute care setting: a compendium of case reports. Pain Manag. 2014;4(4):317-25.
71. Kadiev E, Patel V, Rad P, et al. Role of pharmacogenetics in variable response to drugs: focus on opioids. Expert opinion on drug metabolism & toxicology. 2008;4(1):77-91.
72. Biesiada J, Chidambaran V, Wagner M, et al. Genetic risk signatures of opioid-induced respiratory depression following pediatric tonsillectomy. Pharmacogenomics. 2014;15(14):1749-62.
73. Sadhasivam S, Chidambaran V, Zhang X, et al. Opioid-induced respiratory depression: ABCB1 transporter pharmacogenetics. Pharmacogenomics J. 2015;15(2):119-26.
74. Chidambaran V, Zhang X, Martin LJ, et al. DNA methylation at the mu-1 opioid receptor gene (OPRM1) promoter predicts preoperative, acute, and chronic postsurgical pain after spine fusion. Pharmgenomics Pers Med. 2017;10:157-68.
75. Sadhasivam S, Zhang X, Chidambaran V, et al. Novel associations between FAAH genetic variants and postoperative central opioid-related adverse effects. Pharmacogenomics J. 2015;15(5):436-42.
76. Chidambaran V, Mavi J, Esslinger H, et al. Association of OPRM1 A118G variant with risk of morphine-induced respiratory depression following spine fusion in adolescents. Pharmacogenomics J. 2015;15(3):255-62.
77. Chidambaran V, Pilipenko V, Spruance K, et al. Fatty acid amide hydrolase-morphine interaction influences ventilatory response to hypercapnia and postoperative opioid outcomes in children. Pharmacogenomics. 2017;18(2):143-56.
78. Park HJ, Shinn HK, Ryu SH, Lee HS, Park CS, Kang JH. Genetic polymorphisms in the ABCB1 gene and the effects of fentanyl in Koreans. Clinical pharmacology and therapeutics. 2007;81(4):539-46.

79. Wu WD, Wang Y, Fang YM, Zhou HY. Polymorphism of the mu-opioid receptor gene (OPRM1 118A>G) affects fentanyl-induced analgesia during anesthesia and recovery. Mol Diagn Ther. 2009;13(5):331-7.
80. Henker RA, Lewis A, Dai F, et al. The associations between OPRM 1 and COMT genotypes and postoperative pain, opioid use, and opioid-induced sedation. Biological research for nursing. 2013;15(3):309-17.
81. Romberg RR, Olofsen E, Bijl H, et al. Polymorphism of mu-opioid receptor gene (OPRM1:c.118A>G) does not protect against opioid-induced respiratory depression despite reduced analgesic response. Anesthesiology. 2005;102(3):522-30.
82. Oertel BG, Schmidt R, Schneider A, Geisslinger G, Lotsch J. The mu-opioid receptor gene polymorphism 118A>G depletes alfentanil-induced analgesia and protects against respiratory depression in homozygous carriers. Pharmacogenet Genomics. 2006;16(9):625-36.
83. Zwisler ST, Enggaard TP, Noehr-Jensen L, et al. The antinociceptive effect and adverse drug reactions of oxycodone in human experimental pain in relation to genetic variations in the OPRM1 and ABCB1 genes. Fundam Clin Pharmacol. 2010;24(4):517-24.
84. Somogyi AA, Barratt DT, Coller JK. Pharmacogenetics of opioids. Clinical pharmacology and therapeutics. 2007;81(3):429-44.
85. Kosarac B, Fox AA, Collard CD. Effect of genetic factors on opioid action. Curr Opin Anaesthesiol. 2009;22(4):476-82.
86. Oertel BG, Doehring A, Roskam B, et al. Genetic-epigenetic interaction modulates mu-opioid receptor regulation. Hum Mol Genet. 2012;21(21):4751-60.
87. Ye C, Coco J, Epishova A, et al. A Crowdsourcing Framework for Medical Data Sets. AMIA Jt Summits Transl Sci Proc. 2018;2017:273-80.
88. Callahan A, Fries JA, Re C, et al. Medical device surveillance with electronic health records. NPJ Digit Med. 2019;2:94.
89. Agency for Healthcare Research and Quality. Clinical Classifications Software (CCS) for ICD-9-CM 2017 [Available from: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp.
90. Agency for Healthcare Research and Quality. Clinical Classifications Software (CCS) for ICD-10-PCS (beta version) 2019 [Available from: https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp.
91. SNOMED International. SNOMED-CT: 5-Step Briefing 2021 [Available from: https://www.snomed.org/snomed-ct/five-step-briefing.
92. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.
93. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. Critical care medicine. 2016;44(2):368-74.
94. Roden DM, Pulley JM, Basford MA, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clinical pharmacology and therapeutics. 2008;84(3):362-9.
95. Marees AT, de Kluiver H, Stringer S, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018;27(2):e1608.
96. Reed E, Nunez S, Kulp D, Qian J, Reilly MP, Foulkes AS. A guide to genome-wide association analysis and post-analytic interrogation. Statistics in medicine. 2015;34(28):3769-92.
97. Khanna AK, Bergese SD, Jungquist CR, et al. Prediction of Opioid-Induced Respiratory Depression on Inpatient Wards Using Continuous Capnography and Oximetry: An International Prospective, Observational Trial. Anesthesia and analgesia. 2020;131(4):1012-24.
98. Jungquist CR, Chandola V, Spulecki C, et al. Identifying Patients Experiencing Opioid-Induced Respiratory Depression During Recovery From Anesthesia: The Application of Electronic Monitoring Devices. Worldviews on evidence-based nursing / Sigma Theta Tau International, Honor Society of Nursing. 2019;16(3):186-94.