Preventing Opioid Overdose: From Prediction to Operationalization

by

Neal Kaw

B.A.Sc., University of Toronto (2015)
M.A.Sc., University of Toronto (2017)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Master of Science in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author . . . .

Sloan School of Management
May 7, 2021

Certified by . . . .

Jónas Oddur Jónasson
Class of 1943 Career Development Professor
Assistant Professor, Operations Management
Thesis Supervisor

Certified by . . . .

Nikolaos Trichakis
Zenon Zannetos (1955) Career Development Professor
Associate Professor, Operations Management
Thesis Supervisor

Accepted by . . . .

Georgia Perakis
William F. Pounds Professor of Management
Co-Director, Operations Research Center

Preventing Opioid Overdose: From Prediction to Operationalization

by

Neal Kaw

Submitted to the Sloan School of Management
on May 7, 2021, in partial fulfillment of the
requirements for the degree of
Master of Science in Operations Research

Abstract

The opioid epidemic remains a significant public health challenge in the US. A potential catalyst for reducing the incidence of opioid-related harm is the development and operationalization of risk stratification models. Prior work has focused on the statistical performance of such models without considering operational implications. Predicting the most severe outcome (fatal overdose) is a particular challenge due to imbalanced datasets. We partner with Staten Island Performing Provider System to access claims data and electronic health records for the patient population on Staten Island. For this population, we develop a single machine learning model for predicting a full range of adverse opioid-related events, and achieve an area under the receiver operating characteristic curve of 0.95, 0.87, 0.83 for the outcomes of any adverse opioid event, opioid overdose, and fatal opioid overdose, respectively, even in the absence of training data on fatal overdoses. Subsequently, we conduct a rolling horizon analysis to evaluate the capacity requirements of intervention policies leveraging the model. We find that the model can be used to identify a small intervention cohort (1% of the highest-risk patients) which includes the majority (69%) of adverse opioid events, allowing for targeted interventions with limited intervention capacity. Finally, we quantify the tradeoff between predictive performance and concerns that arise in implementation, such as interpretability, delay in data feeds, and prediction window length. Our results suggest that predictive performance does not need to be sacrificed to satisfy implementation concerns.

Thesis Supervisor: Jónas Oddur Jónasson
Title: Class of 1943 Career Development Professor
Assistant Professor, Operations Management

Thesis Supervisor: Nikolaos Trichakis
Title: Zenon Zannetos (1955) Career Development Professor
Associate Professor, Operations Management

Acknowledgments

I am grateful to my advisors, Jónas Oddur Jónasson and Nikos Trichakis, for being outstanding mentors in research, and for being exceedingly kind, generous, and patient, especially in the face of challenges. I have also been fortunate to collaborate with Deeksha Sinha on the research in this thesis, and I thank her for being a dedicated and hardworking project partner.

I would like to thank our collaborators from SI PPS, Anyi Chen, Joseph Conte, Dhruvit Patel, Ashley Restaino, Mark Slavutsky, and Salvatore Volpe, for making this research possible and providing valuable input that shaped its development. I am also thankful to Anne Quaadgras for facilitating this collaboration and providing insightful feedback.

I thank my teachers and colleagues from Toronto for always being available for advice and support: Tim, Mike C., Roy, Ali, Mike F., Michaelia, Josh, David, Muhammad. Finally, I thank my friends and peers in the ORC, especially Evan, Shuvo, Kevin, Matt, Arthur, Emma, and Kayla, for helping me on so many occasions, and for all the memorable experiences.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

Opioid overdoses and opioid use disorder are significant public health challenges in the United States: in 2018, 47,600 drug overdose deaths involved an opioid (Wilson et al., 2020), an estimated 2 million people had an opioid use disorder, and over 10 million people misused opioids (Substance Abuse and Mental Health Services Administration, 2019). Battling an epidemic of such magnitude requires persistent, coordinated, and multi-pronged efforts, ranging from prevention to prophylaxis to treatment. A potential catalyst for reducing the incidence of opioid-related harm is the effective risk stratification of individual patients, as identification of high-risk patients is the foundation for successful preventive actions, such as targeted interventions.[1] In this thesis, we investigate how to build and operationalize predictive models for opioid risk stratification.

Imagine a healthcare organization that provides care to a patient population susceptible to opioid-related harm. The organization aspires to build a risk stratification model that would identify patients at high risk of adverse events. The model, which could either be a black-box machine learning model or a more interpretable one, would assign personalized risk scores by predicting the risk of individuals for some adverse outcome such as abuse, poisoning, or even death due to opioids. To make accurate predictions, the model would use available electronic health record (EHR)

[1] Indeed, multiple works suggest that early detection of opioid abuse can help prevent addiction (Fareed et al., 2011; Fishman et al., 2000).

and prescription data, some of which might not be up-to-date, depending on data feed access. The organization plans to employ some number of social workers, counsellors, or psychologists, who could provide additional preventive care to patients at risk. Operationally, the organization would then periodically apply the model to identify the highest risk patients and subsequently deploy these scarce and costly human resources to provide targeted interventions.

The above is a prototypical and representative application of opioid risk stratification models that we have envisioned together with our partners from Staten Island Performing Provider System (SI PPS), an alliance of clinical and social service providers on Staten Island in New York City. By outlining the operationalization of such a system, we have identified five statistical and operational challenges, detailed below, believed by SI PPS to be the most important ones to address prior to implementation. Each challenge is likely to be common across any practical application of opioid risk prediction models. Our research strives to address these challenges and our aspiration is that this work will facilitate widespread adoption of risk stratification models that could help curb the opioid epidemic. While prior work has proposed machine learning models to predict a patient's risk of adverse opioid outcomes, it has been mostly statistical in nature. To our knowledge, no rigorous investigation of the implementation challenges, discussed below, has been conducted.

The first challenge (§2.4) is of a statistical nature and relates to using the appropriate outcome variable. In particular, depending on the use case, the models could be calibrated to predict less severe outcomes, such as a diagnosis of opioid abuse, or more severe outcomes, such as opioid-related poisoning or death. The latter outcomes remain rare enough that they result in highly imbalanced datasets, making it a a challenge to build an effective prediction model. Indeed, as we discuss in our literature review (§1.3), only a few prior studies have presented models built to predict fatal overdoses (often using standard oversampling techniques), with modest success.

The second challenge (§3.1) relates to the fundamental operational problem of capacity planning. Given an accurate model to stratify patients into risk categories, what are the provider organization's resource requirements as it plans to selectively intervene to mitigate a targeted percentage of adverse outcomes? For example, how many patients would need to be preemptively enrolled in an intervention program to help avert 50% of opioid overdoses? This is an important operational question, since a larger intervention cohort would clearly increase the number of future adverse outcomes identified correctly, but would also increase the required resources for intervention, and this relationship may show a pattern of diminishing returns.

The other three challenges relate to design implementation choices involving tradeoffs that could compromise prediction accuracy. In particular, the third challenge deals with interpretability of risk stratification models, which is often valued by implementing organizations (§3.2.1). Interpretable models can improve the face validity of the predictions for healthcare providers, and also aid in designing appropriate interventions. However, the key question is to what extent, if any, this preference affects the attainable prediction accuracy. The fourth challenge (§3.2.2) is to select the appropriate prediction window. We expect that the length of a prediction window affects model accuracy, but the best choice from the perspective of prediction accuracy may not be appropriate for a particular intervention. Similarly, the fifth challenge (§3.2.3) is to develop accurate prediction models, even for settings where the required input data may not be up-to-date. Understanding how data delay affects prediction accuracy allows an implementing organization to appropriately prioritize investment in IT infrastructure to ensure the availability of up-to-date data.

1.1 Overview of methods and results

We address the implementation challenges described above in partnership with SI PPS. In 2018, Staten Island had a rate of 28.7 unintentional opioid overdose deaths per 100,000 population, which was the highest rate among the five boroughs of New York City (New York City Department of Health and Mental Hygiene, 2019), and nearly double the rate of 14.6 for the US overall (Wilson et al., 2020). Through our partnership with SI PPS, we access patient data from the Medicaid-insured population residing on Staten Island and the uninsured population receiving treatment on Staten Island, a total of 251,781 patients. Our dataset (§2.1) links this population's medical and pharmacy claims; EHR data from both of the hospitals and three of the four federally qualified health centers (FQHCs) on Staten Island; and opioid overdose deaths recorded by the office of the Richmond County District Attorney. This dataset allows us to construct predictor variables capturing information on filled prescriptions, disease diagnoses, encounters with providers, and demographics, all of which are among the known risk factors for opioid-related harm. We construct three nested outcome variables of increasing severity: *AnyOpioidHarm* corresponds to the occurrence of any opioid overdose, dependence, abuse, or adverse effects, or medication-assisted treatment (MAT) for opioid use disorder (OUD); *Overdose* corresponds to an opioid poisoning; and *FatalOverdose* corresponds to a fatal opioid poisoning.

We summarize three main findings. First, we utilize a novel approach to address data imbalance due to the low frequency of *FatalOverdose* observations. To this end, we leverage the nested structure of our outcomes and use the least severe adverse event (*AnyOpioidHarm*) to train risk stratification models for more severe outcomes (*Overdose* and *FatalOverdose*). This approach resembles oversampling or synthetic generation of positive samples, both of which have been suggested in the machine learning literature as methods to deal with imbalanced datasets. However, our tailored approach is superior to the aforementioned generic methods, because the new positive samples we add are drawn from distinct and genuine patients, as opposed to duplicate or synthetically generated ones. Further, these samples also correspond to patients who experienced adverse outcomes that are often precursors to the more severe outcomes. With a 90-day prediction window, we find that the model trained to predict *AnyOpioidHarm* has stronger performance in predicting *Overdose* (AUC of 0.87) or *FatalOverdose* (AUC of 0.83) than the models specifically trained for predicting those outcomes, respectively.

Our novel approach also leads to a single parsimonious model that can be used to predict different outcomes. What's more, the predictive power of this parsimonious model compares favorably with the best results in the existing literature, for all outcomes. This means that even if providers have different interventions tailored for outcomes of different severity, only a single predictive model may be needed rather than different ones tailored for each outcome, affording simplicity, easier implementation and maintenance. Moreover, *FatalOverdose* is difficult to predict, in general, because of the relative rarity of the outcome, and is often infeasible to predict because cause of death is generally not recorded in the claims and EHR databases used to fit these models. In this scenario, our finding shows that a model which predicts a less severe outcome would suffice to guide an intervention targeted to prevent *FatalOverdose*.

Second, we find that intervening with a small fraction of patients can identify and potentially prevent a large fraction of opioid-related harms. Since the number of patients whose outcomes could be prevented depends not only on the accuracy of the predictive model, but also on the effectiveness of the intervention, we conduct two separate analyses: one in which adverse events can only be "caught" if patients are actively receiving an intervention, and one in which all future adverse events for given patient are "caught" if the patient has received the intervention once. These analyses provide upper and lower bounds on the intervention capacity required to prevent a given percentage of adverse events. We find that in order to help avert 50% of instances of *AnyOpioidHarm*, *Overdose*, or *FatalOverdose*, somewhere between 0.3-0.5%, 1.2-2.4%, and 0.9-3.5% of patients need to receive intervention, respectively, when their selection is guided by our risk stratification models.

Third, our results suggest that predictive performance does not need to be sacrificed to navigate the design implementation tradeoffs we consider. With regards to interpretability, we fit an optimal classification tree (OCT) model (Bertsimas and Dunn, 2017), which is highly interpretable because it consists of only a single decision tree. We find little difference in AUC compared to XGBoost (eXtreme Gradient Boosting, Chen and Guestrin (2016)) models, but great improvement in interpretability, given the opacity of an ensemble of trees. Similarly, we also find relatively little variation in performance when we fit models to predict risk over different durations of time in the future, meaning the duration can be chosen to suit the intervention of interest without concern for performance. Finally, we find that prediction performance is not significantly affected when there is a delay in populating the underlying dataset, suggesting that investments in IT infrastructure to reduce this delay may not be a prerequisite for effective risk stratification.

1.2 Contributions

We make three main academic contributions, beyond the results specific to our study setting. First, this study is the first to show that opioid overdoses, fatal or non-fatal, can be predicted accurately using a model that was trained to predict *any* opioid harm. Second, this study is the first to estimate the capacity requirements for an intervention that targets patients on the basis of an opioid risk stratification model. Existing works on early warning systems for opioid-related harm only report the prediction accuracy of the resulting models, without considering the operational resources that would be necessary to identify or prevent a targeted number of outcomes using the model. Third, our research is the first to systematically explore how the performance of a risk stratification model for opioid-related harm is affected by the tradeoffs that arise in the operational deployment of the model. Prior research develops predictive models for different settings: for example, Hastings et al. (2020) predict risk of opioid-related harm in the next 5 years, for a patient who has never experienced such harm, whereas Lo-Ciganic et al. (2019) predict risk of overdose in the next 3 months, for a patient receiving prescription opioids and who may have previously experienced opioid-related harm. Comparing the performance of models for these different settings is challenging because of differences in populations and variable definitions. Our study allows such comparison by using the same dataset to fit models for outcomes of different severity, for interpretability, for different prediction windows, and for different delays in populating the underlying database.

1.3 Literature review

We review two streams of literature related to this thesis: predictive models that estimate risk of adverse opioid-related outcomes, and the operational deployment of predictive models.

1.3.1 Predictive models for opioid-related harm

Clinical guidelines for the prescription of opioids for chronic pain recommend that clinicians should estimate a patient's risk for opioid-related harm, in particular for overdose (Chou et al., 2009; Dowell et al., 2016; Nuckols et al., 2014). Accordingly, a large body of literature applies traditional statistical techniques to infer which clinical and demographic characteristics of patients are associated with higher risk of opioid-related harm. Reviewing many of these papers, Webster (2017) and Cragg et al. (2019) find that the risk factors for opioid misuse or addiction include current or previous substance use or abuse, mental health disorders, younger age, and male sex. Opioid abuse or dependence is also associated with higher numbers of outpatient, inpatient, or emergency department (ED) encounters; opioid prescriptions with higher days supply, or a greater number of opioid classes; and higher numbers of concurrent prescriptions, including those for antidepressants and benzodiazepines (Cochran et al., 2014). Reviews of risk factors for *Overdose* (Park et al., 2016; Webster et al., 2011) show an association with mental health comorbidities, substance use, and use of long-acting opioids[2]. *FatalOverdose* is also associated with number of filled prescriptions, prescriptions of specific types of opioids (Paulozzi et al., 2012), use of muscle relaxants, and use of Schedule II opioids[3] (Garg et al., 2017).

Motivated by these studies of risk factors for opioid-related harm, several papers propose tools for risk assessment of patients, including questionnaire-based screening tools, regression models, and machine learning models. Two screening tools that as-

[2] Long-acting opioids are drug formulations that have a longer-lasting pain-relieving effect, as compared to short-acting opioids (Argoff and Silvershein, 2009).

[3] Drugs are classified into schedules (I to V) by the US Drug Enforcement Administration (DEA), where increasing schedule corresponds to higher potential for abuse, and Schedule I drugs are defined such that they have no accepted medical use (United States Drug Enforcement Administration, n.d.).

sign a risk score to a patient based on their self-reported answers to a questionnaire are the Screener and Opioid Assessment for Patients with Pain (SOAPP) (Butler et al., 2004) and the Opioid Risk Tool (ORT) (Webster and Webster, 2005). Although these tools incorporate questions relating to several of the risk factors identified above, they have limited accuracy (Dowell et al., 2016). More sophisticated approaches use logistic regression or Cox proportional hazards models to predict various outcomes, including opioid overdose (Chang et al., 2019; Zedler et al., 2015, 2018), opioid or heroin overdose (Glanz et al., 2018), fatal opioid overdose (Ferris et al., 2019; Saloner et al., 2020), opioid overdose resulting in death or hospitalization (Geissert et al., 2018), nonfatal opioid overdose (Saloner et al., 2020), and opioid misuse, opioid abuse, or OUD (Dufour et al., 2014; Hylan et al., 2015; Rice et al., 2012; Tarter et al., 2020; White et al., 2009). Going beyond these regression-based approaches, several recent papers propose using random forests, gradient boosting machines, or deep neural networks to predict opioid overdoses (Dong et al., 2019; Ellis et al., 2019; Lo-Ciganic et al., 2019) and incidence of OUD (Hasan et al., 2019; Lo-Ciganic et al., 2020). Hastings et al. (2020) additionally use recurrent neural networks to predict the occurrence of opioid abuse, dependence, or poisoning. The works reviewed here generally construct predictor variables corresponding to known risk factors, using data sourced from insurance claims, EHRs, or state-level prescription drug monitoring programs. However, Saloner et al. (2020) and Hastings et al. (2020) additionally incorporate criminal justice data, and Hastings et al. (2020) also use data on employment and social benefits.

The preceding articles focus primarily on using standard methods to develop predictive models for a single outcome of interest, motivated by a single intervention setting. Our study goes beyond standard approaches to propose a novel solution to the class imbalance problem. Furthermore, our study is the first to systematically explore the prediction of a full range of opioid-related harms, and tradeoffs between predictive performance and interpretability, prediction window length, and data delay. More generally, we study the performance of these models under operational deployment, which is the next topic we review.

1.3.2 Operationalizing predictive models

Several papers consider queuing models and study the effect of admission control policies that incorporate some predicted information on customer arrivals. Yom-Tov et al. (2020) study a setting where we have a predictive model that stratifies prospective arrivals into classes corresponding to the expected revenue that would be received if service were provided. They consider the question of how many customers of each class to invite into the queuing system, and when, to maximize expected revenue. The next three papers suppose that we have a model which predicts customer arrivals in advance. Under this setting, Delana et al. (2021) and Zhang et al. (2016) consider the policy of proactive service, i.e., using an opportunity when servers are idle to proactively serve a customer who has not yet arrived. Delana et al. (2021) find that this policy reduces wait time, even if predictions have limited accuracy and not all identified customers are actually willing to be served proactively. Zhang et al. (2016) show that the average wait decreases exponentially in the length of the future window for which we have predicted information. Xu and Chan (2016) specifically propose policies for when to divert patient arrivals away from an emergency department (ED), and show using simulation that their policies reduce ED wait times while serving the same number of patients.

Another group of papers uses more diverse methods to study how to optimally use a predictive model to decide whether to transfer a patient to or from an intensive care unit (ICU). In the former case, the decision maker has a model that predicts for each individual patient in an inpatient ward their risk of deterioration requiring transfer to an ICU. To potentially reduce their length of stay or improve their eventual outcome, we can proactively transfer a patient at high estimated risk, with the possible downside that we waste ICU capacity on a patient who was wrongly identified, and take capacity away from a patient who needs ICU capacity urgently (rather than one who is only predicted to need ICU capacity.) Hu et al. (2018) provide empirical evidence of the benefit of proactive transfer to ICU on the basis of a predictive model, and also use a simulation model to evaluate three proposed policies.

Hu et al. (forthcoming) propose a queuing model to describe this setting and derive a cost-minimizing scheduling policy for transfers to ICU. Grand-Clément et al. (2020) propose a robust Markov decision process (MDP), where each patient's predicted risk of deterioration is incorporated into the state space, and they derive a policy that minimizes in-hospital mortality and length-of-stay. Finally, Cheng et al. (2019) also propose an MDP incorporating each patient's prediction in the state space, but they instead consider the question of when to *stop* treatment in ICU, given the costs of early and delayed ending of treatment.

We finally mention two additional applications related to capacity planning. Peck et al. (2012) develop models that predict, upon triage at an ED, the probability of whether a patient will be admitted to hospital. They then use the predictions for all patients currently in the ED to estimate the total number of hospital beds that will be required, which is intended to help bed managers plan in advance. Kurz and Pibernik (2016) are motivated by a setting where a provider of maintenance services for aircraft engines is able to predict in advance the number of customer arrivals for a future period of time. They then study the question of when add to an M/M/1 queuing system extra capacity which can only be used for a limited amount of time, to reduce average wait time.

Our study contributes to this stream of literature by studying novel aspects of operationalizing a predictive model. The bulk of the preceding articles study how predictive models can be used in admission control policies for a service system to achieve objectives such as maximizing revenue, minimizing cost, or minimizing wait time. On the other hand, we investigate the intervention capacity that is required to prevent a targeted percentage of the predicted events; Peck et al. (2012) and Kurz and Pibernik (2016) similarly consider capacity questions, but for different objectives. We also consider how the quality of predictions is affected when the prediction task is tailored to the intervention, in terms of interpretability and prediction window length, and under the operational constraint of data delay.

1.4 Organization

The body of this thesis is divided into two chapters, aligned with the two streams of literature we have reviewed. In Chapter 2, we describe our data sources and our machine learning methodology, including our novel approach to the class imbalance problem, and we benchmark our model's predictive performance against comparable works in the literature. Chapter 3 has two parts: first we estimate the intervention capacity required to use our model to prevent a targeted percentage of adverse outcomes, and second we quantify the tradeoffs between predictive performance and interpretability, prediction window length, and data delay. In Chapter 4, we offer concluding remarks.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

Opioid-related risk stratification

This chapter describes our approach to building prediction tools for stratifying patients based on their risk for different opioid-related adverse outcomes. That is, we seek to build models that assign risk scores to individuals, with the goal that patients who would experience an adverse event are more likely to be assigned a higher risk score ex ante, relative to those who would not experience an adverse event. We first describe our data in §2.1-2.3 and then our methodology in §2.4. An important innovation of our models is how we deal with dataset imbalance stemming from the excessively low incidence rates of severe outcomes. In a nutshell, because all extant generic methods to deal with data imbalance failed in our study, we propose a tailored approach that leverages the nested structure of the outcomes we consider. We discuss this approach in detail in §2.4.1. We report the test set performance of our models in §2.5, and then discuss how the performance compares against prior results reported in the literature in §2.6.

2.1 Data sources

We partner with SI PPS to access several databases containing data for both Medicaid-insured and uninsured patients. For the Medicaid-insured population that has an address of residence on Staten Island, the data includes adjudicated pharmacy and medical insurance claims spanning July 2014 to February 2020, provided by the New

York State Department of Health. For these patients, the data includes linked EHRs spanning the same time period from both hospitals on Staten Island, and three of the four FQHCs, which provide ambulatory care. These EHRs also include data on care provided to uninsured patients at either hospital, whether or not they resided on Staten Island. Finally, we also link a dataset from the Richmond County District Attorney, which records instances of death due to opioid overdose from January 2018 to February 2020, within the boundaries of Staten Island.

All data analysis was performed in a cloud-based data warehouse with extensive security measures, validated by external audits, to protect data from unauthorized access. All data was deidentified to prevent attribution to any individual.

2.2 Dataset construction

Our study population includes all patients with any encounter or filled prescription in the period spanned by the dataset. In this context, an encounter refers to any inpatient or ED visit at either of the Staten Island hospitals, or any outpatient visit at any of the three FQHCs covered by our data. A filled prescription refers to any prescription which has been dispensed to the patient by an outpatient pharmacy. Each patient enters the cohort on the date of their first encounter or filled prescription, and then remains in the cohort until end of observation, cancer diagnosis, or death (whichever comes first). We study a broad population, rather than restricting ourselves to a population such as patients using prescription opioids (e.g., Lo-Ciganic et al. (2019)), because opioid-related harms may also be caused by illicit rather than just prescribed opioid use. The motivation for the cancer exclusion is that the aggressive use of opioids for treating pain in patients with cancer is strongly supported by the medical community despite the associated risks (National Academies of Sciences, Engineering, and Medicine, 2017), and therefore a risk prediction tool developed for the general population would not be appropriate for this subpopulation. The exclusion is implemented using International Classification of Diseases (ICD) codes (Table A.1 in Appendix A), and is also common in other studies examining prediction of

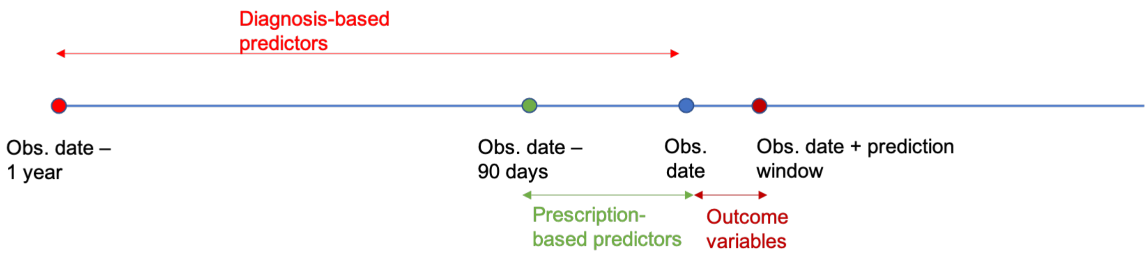

Diagnosis-based predictors
Obs. date – 1 year
Obs. date – 90 days
Obs. date
Obs. date + prediction window
Prescription-based predictors
Outcome variables

Figure 2-1: Construction of an observation in the dataset

opioid-related disorders, e.g., Lo-Ciganic et al. (2019, 2020).

The dataset we construct for analysis has columns consisting of predictor and outcome variables, and rows which we refer to as observations. An observation corresponds to a specific patient at a specific point in time, capturing the patient history prior to this time point in the form of predictor variables, and capturing future events in the outcome variables *AnyOpioidHarm*, *Overdose*, and *FatalOverdose*. Specifically, an observation records the values of diagnostic-based predictors using data from the preceding year; prescription- and encounter-based predictors using data from the preceding 90 days; and the occurrence or non-occurrence of each possible outcome using data from a period following the observation date (the *prediction window*, for which we will test multiple values in §3.2.2). See Figure 2-1 for an illustration. More detail on the predictor and outcome variables is given in the following two subsections.

To construct observations, we first identify the earliest chronological appearance of any patient in the data. For each patient, we then record an initial observation one year after that date, and every 90 days thereafter. We repeat the above-described procedure twice, to construct one dataset for the *FatalOverdose* outcome and another, distinct dataset, for the other two outcomes. This is because the data for the *FatalOverdose* outcome is available only from January 2018 to February 2020, which is a subset of the time period spanned by the remaining data sources, and therefore we only use this subset of the data to construct a corresponding dataset.

2.2.1 Predictor variables

For each observation, we construct a total of 107 predictor variables (Tables A.2-A.4 in Appendix A), motivated by the known risk factors for opioid-related harm identified in the literature reviewed in §1.3. Of the predictors, 73 are related to filled prescriptions in the past 90 days. For each long-acting or short-acting formulation of each prescription opioid compound, we count the number of prescription fills and the total days supplied; an example variable is the number of prescriptions of short-acting hydrocodone filled in the past 90 days. For each of antidepressants, muscle relaxants, benzodiazepines, gabapentinoids, and pregabalin we also count the number of prescription fills and the total days supplied; an example variable is the total days supply of antidepressant prescriptions filled in the past 90 days. We also count the number of days of co-prescription of benzodiazepines and opioids. Finally, for each of the five drug schedules, we construct an indicator variable for whether an opioid prescription in that schedule has been filled. Drugs are uniquely identified by their FDA-issued National Drug Code (NDC): sources for NDCs for non-opioid drugs are in Table A.5 in Appendix A, and NDC codes and DEA schedule for opioids are published by the National Center for Injury Prevention and Control (2018).

Twenty-three of the predictors are related to diagnostic codes entered in the past year for mental health disorders, alcohol use disorders, and (non-opioid) drug use disorders. Specifically, we count the number of diagnostic codes entered for anxiety, depression, schizophrenia, other psychosis, bipolar disorder, attention deficit hyperactivity disorder (ADHD), and personality disorders. For non-opioid drugs and for alcohol, we count the number of diagnostic codes entered for use, adverse effects, abuse, dependence, poisoning, or other related diagnoses. ICD codes for these disorders are found in the Chronic Condition Data Warehouse (2019).

Eleven variables remain. We identify each patient's gender, race and ethnicity. We compute the patient's age on the observation date, and we also count their number of inpatient, outpatient, and ED encounters in the past 90 days. Finally, we also count the number of previous observations for the same patient that had a positive value for each of four outcome-based binary indicators, which will be defined in the following subsection: *Overdose*; *Overdose* or opioid dependence; *Overdose*, opioid dependence, opioid abuse, or filled prescriptions for medication-assisted treatment (MAT) for OUD; and *AnyOpioidHarm*.

2.2.2 Outcome variables

For the prediction window of a given observation, we identify all occurrences of the following nine possible indications of opioid-related harm: an instance of death due to opioid poisoning; diagnostic codes for opioid poisoning, dependence, abuse, adverse effects, or unspecified use; and filled prescriptions with NDC codes for buprenorphine, methadone, or naltrexone. The latter three drugs are used in MAT for OUD, and the occurrence of these prescriptions should therefore identify any patient who developed OUD, but for whom a diagnosis code is missing in the data. Buprenorphine is itself an opioid which can be used either for pain relief or MAT for OUD (National Center for Injury Prevention and Control, 2018), so in our definition we only include drug formulations used in MAT for OUD. For these nine types of opioid-related harm, Table A.5 and National Center for Injury Prevention and Control (2018) list NDCs, and Table A.6 lists ICD codes.

We then use the nine types of events to construct three nested binary outcome variables of increasing severity. The variable *AnyOpioidHarm* corresponds to the occurrence of any of the nine indicators. The variable *Overdose* corresponds to an opioid poisoning, and *FatalOverdose* corresponds to a fatal opioid poisoning.

2.3 Data limitations

Three potential limitations of our dataset are censoring of variables, historical interventions that prevented adverse outcomes, and a biased representation of the Staten Island population.

In our data, both predictor and outcome variables may be potentially censored. Predictor variables may be censored if, for example, a patient visits an outpatient clinic which is not located on Staten Island and this visit is therefore unaccounted for in the variable which counts the number of outpatient visits. This censoring could affect predictive power. In practice, however, data environments in which predictive models would actually be deployed are very likely to exhibit the exact same form of censoring. Therefore, our analysis, being conducted under circumstances that mimic practice, is likely to reveal the effectiveness of such predictive models when implemented. Outcome variables may also be censored if, for example, a patient died of an opioid overdose outside the boundaries of Staten Island, and the death therefore would not be recorded in our data sourced from the District Attorney. In this case, an observation that ought to be labelled as positive would be labelled as negative. Although this remains a limitation, it may be mitigated by the fact that our population resides on an island, potentially reducing the extent of travel outside its boundaries, and also the number of outcomes of interest that occur outside its boundaries.

Prior interventions may have altered the outcomes that occurred for patients in the past. This would pose a challenge to fitting predictive models, since interventions that had a material effect (for example, preventing an outcome that would have otherwise occurred) would alter the relationship between predictor and outcome variables. However, all previous intervention efforts that SI PPS has been involved in have targeted patients who had already had an overdose, and therefore the number of patients with altered outcomes would be limited. There have not been any prior interventions that proactively targeted at-risk individuals. In fact, our study is a first step towards designing and managing such interventions that SI PPS aspires to deploy.

Finally, there is a possibility that our data contains a biased representation of the Staten Island population. For example, our data only includes records from three of the four FQHCs on Staten Island, and the missing FQHC only serves patients who are developmentally disabled, so these patients are likely underrepresented in our dataset. However, this form of bias is likely not to affect our analysis because the predictive models that we develop would only be applied to exactly the population of patients that is represented in the dataset.

2.4 Methodology

For each of our two datasets (one for the *FatalOverdose* outcome and another for the other two outcomes, as described in §2.2), we divide the observations into two sets: training set (spanning first 70% of the data duration) and testing set (spanning last 30% of the data duration) based on the observation date. We perform K-fold cross-validation on the training set ($K = 3$ for the *FatalOverdose* outcome, and $K = 5$ for the other two). All performance metrics are reported based on the performance of the models on the testing set.

Summary statistics of these datasets are provided in Table 2.1. The non-death dataset that spans a larger time duration has approximately 2.4 and 1.2 million observations in the training and testing sets, respectively. The death dataset, spanning slightly over two years, has approximately 1.2 and 0.7 million observations in the training and testing sets. The number of patients is roughly the same in the training and testing sets in each of the two datasets (close to 0.25 million). Notably, the outcomes in such datasets are severely imbalanced, with very low incidence rates, particularly for the *FatalOverdose* outcome, as also reported in antecedent studies. We discuss this important issue and mitigating actions in §2.4.1.

We train models for predicting the three outcomes of interest over a 90-day prediction window. We train our prediction models using a scalable gradient boosting framework, XGBoost (Chen and Guestrin, 2016). XGBoost is a machine learning method in which an ensemble of decision trees is fit sequentially, and each successive tree is trained to correct the errors made by previous trees. Given the large size of our datasets, the scalable nature of the XGBoost framework makes it particularly attractive for our application.

Algorithms for regression, classification, and ranking tasks are available in the XGBoost framework. While predicting each patient's risk of an adverse opioid-related outcome can be modeled as a classification task, our objective is to stratify patients

Table 2.1: Summary statistics of the datasets

	Death Dataset		Non-death Dataset	
	Training	Testing	Training	Testing
Dataset:				
First observation date	12/31/17	3/26/19	6/30/15	9/12/18
Last observation date	12/26/18	9/22/19	6/14/18	9/7/19
No. of observations	1,165,107	733,828	2,408,815	1,211,499
No of. patients	247,497	249,229	246,377	251,781
Outcomes:				
No. of *AnyOpioidHarm* events			27,084 (1.12%)	7,577 (0.62%)
No. of *Overdose* events			1,698 (0.07%)	446 (0.03%)
No. of *FatalOverdose* events	49 (0.004%)	18 (0.002%)		

Notes: For the testing data, statistics based on a 90-day prediction window are reported. The outcomes are ordered in increasing order of severity.

into categories of higher- and lower-risk. Therefore, we view our task as a bipartite ranking problem (Freund et al., 2003) where the goal in training is to fit a model which ranks patients who had an opioid-related outcome higher than those who did not. We therefore use a ranking objective for model training. As a secondary analysis, we include results from a model trained using a classification objective in Table B.1 in Appendix B.

2.4.1 Dataset imbalance

The datasets are highly imbalanced in terms of the ratio of positive observations (observations with an opioid-related outcome within the prediction window) to negative observations (observations with no opioid-related outcome within the prediction window). Further, there is an order of magnitude reduction in the fraction of positive observations as the severity of outcome increase, i.e., when we compare *AnyOpioidHarm* to *Overdose* and *Overdose* to *FatalOverdose*. Specifically, for the least severe outcome (*AnyOpioidHarm*), 0.6 - 1.1% of observations are positive, whereas for the most severe outcome (*FatalOverdose*), 0.002% - 0.004% of observations are positive (Table 2.1). This makes prediction challenging, especially for the more severe outcomes of *Overdose* and *FatalOverdose*.

In the statistics and machine learning literature, various heuristics have been proposed to deal with imbalanced datasets for prediction tasks (Sun et al., 2009). Key approaches are oversampling the minority class, undersampling the majority class, or introducing new samples of the minority class before training the model. New samples are typically generated using existing samples from the minority class. For example, Lee (1999) proposes taking samples from the minority class and adding Gaussian noise to generate new samples, Menardi and Torelli (2014) propose taking samples from the minority class and generating new samples in its neighborhood as specified by a probability distribution, and Chawla et al. (2002) propose generating new samples as those lying on a line connecting a sample from the minority class with its nearest neighbor.

Application of the aforementioned heuristics to deal with the dataset imbalance failed in our study. This is consistent with prior literature showing that the effectiveness of oversampling is limited for high-dimensional data (Blagus and Lusa, 2012). In our experiments, model performance did not improve after adding synthetically generated observations, although there was a significant increase in the algorithm's runtime, due to the increased dataset size.

We propose a novel way to deal with this challenge within the context of our problem. As less severe outcomes in our setting are often precursors to more severe outcomes, positive observations from the less severe outcome make for good synthetic positive observations when training a model for the more severe outcome. This proposal adapts the core idea of the standard methods, which is to generate new samples close to existing positive samples. Our approach is superior to oversampling or creating new samples as it does not overfit to the existing positive samples, and the new positive samples are also true patient samples. Thus, when training models for predicting *Overdose* and *FatalOverdose*, we treat all observations with a positive *AnyOpioidHarm* outcome as positive observations. Due to the nested structure of our outcomes, this implies using a single model trained on the *AnyOpioidHarm* outcome to predict all three outcomes, leading to a *parsimonious* modeling approach.

2.4.2 Models trained and performance evaluation

In summary, we train three XGBoost models, one to predict risk for each of the three outcomes. The model trained on the *AnyOpioidHarm* outcome also serves as a parsimonious model that predicts risk for all outcomes, following the idea proposed in §2.4.1.

A standard way to evaluate the performance of prediction models for risk stratification is through the receiver operating characteristic (ROC) curve. The ROC curve plots the sensitivity (or recall), the fraction of positive outcomes that the model identifies correctly, against its specificity, the fraction of negative outcomes that the model identifies correctly. Ideally, a prediction model should have both high sensitivity and high specificity. A commonly used summary statistic is the area under the curve. In particular, the area under the ROC curve (AUC) equals the probability that if one positive and one negative observation are chosen randomly, the model will identify the positive observation as higher-risk. This metric is aligned with our goal of ranking patients who would experience an adverse outcome as higher risk than those who would not.

2.5 Results

We first discuss results for the three models separately trained for each of the three outcomes. The corresponding AUC results are presented in Table 2.2. The AUC obtained by the model predicting *AnyOpioidHarm* is highest at 0.95. This is followed by the model for predicting *Overdose* (with an AUC of 0.87). The least AUC (0.68) among these three outcomes is obtained for the *FatalOverdose* outcome. This is not surprising given the extremely low incidence of the *FatalOverdose* outcome, leading to a highly imbalanced dataset that makes learning challenging.

We now discuss the performance of the parsimonious model, which we introduced as a way to mitigate the imbalance that seems to compromise AUC for the *FatalOverdose* outcome in particular. Our results are reported in Table 2.3. For the *Overdose* outcome, given that the model trained directly on that outcome already

Table 2.2: AUC of XGBoost models for different outcomes

Outcome	AUC
AnyOpioidHarm	0.95 (0.95 - 0.95)
Overdose	0.87 (0.85 - 0.89)
FatalOverdose	0.68 (0.54 - 0.82)

Notes: Parentheses report 95% confidence intervals.

Table 2.3: AUC of parsimonious XGBoost model for different outcomes

Outcome	AUC
AnyOpioidHarm	0.95 (0.95 - 0.95)
Overdose	0.87 (0.85 - 0.89)
FatalOverdose	0.83 (0.73 - 0.92)

Notes: Parentheses report 95% confidence intervals.

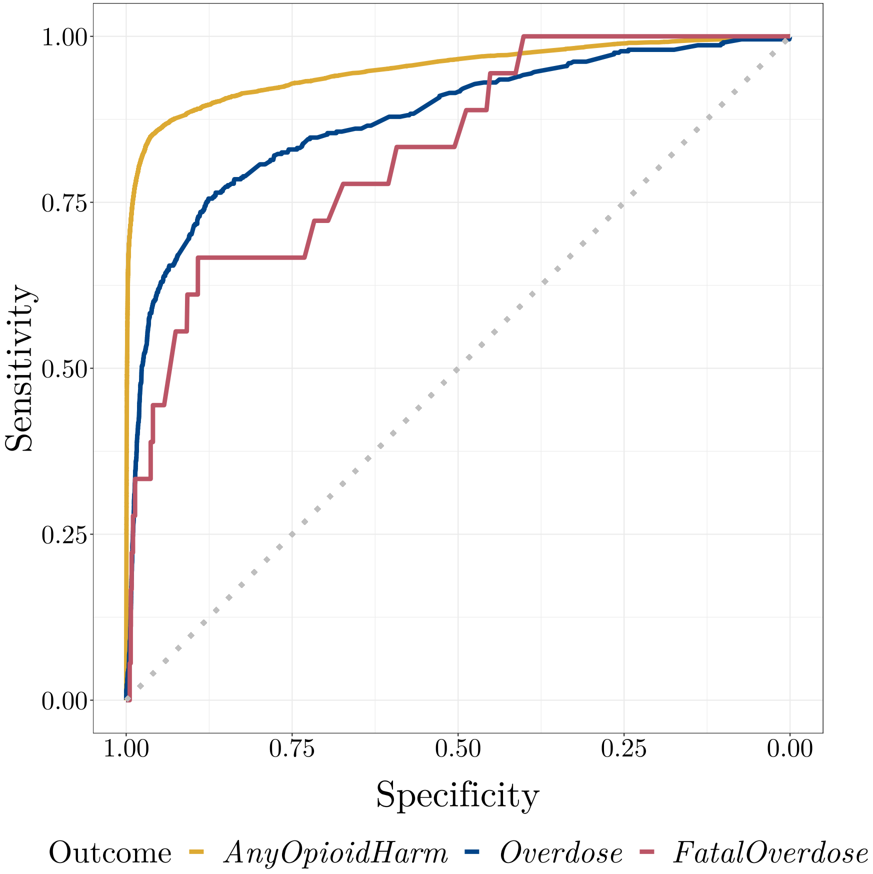
enjoyed strong performance, it is not surprising that we observe only a modest improvement in AUC (beyond the second decimal point) when using the parsimonious model. However, for the *FatalOverdose* outcome, AUC increases from 0.68 to 0.83. Thus, the parsimonious modeling approach effectively deals with the challenge of data imbalance for the *FatalOverdose* outcome.

To understand the parsimonious model's sensitivity at different levels of specificity, we plot the ROC curves for the three outcomes in Figure 2-2. For all three outcomes, we observe performance much better than random guessing (denoted by the light gray line in the figure). In particular, for the model predicting *AnyOpioidHarm*, we observe that the ROC curve reaches very closely to the upper-left corner, indicating simultaneously achieved high sensitivity and specificity.

2.6 Comparison with prior literature

How well does our parsimonious model perform compared to those presented in the literature? The differences in study design, such as population under study, prediction

Figure 2-2: ROC curves for the parsimonious XGBoost model



window, and the exact definition of the opioid outcome, make a direct comparison with other works challenging. When adjusting for such differences, our model approximately matches or outperforms existing models in terms of AUC by a few percentage points. We provide details for each of the three outcomes.

For the *AnyOpioidHarm* outcome, Hasan et al. (2019) consider a similar setting to ours and report an AUC of 0.97, which our model approximately matches. For the *Overdose* outcome, Dong et al. (2019) undersample the negative observations so that at least 9% of observations are positive, whereas Lo-Ciganic et al. (2019) and Geissert et al. (2018) limit the study population to patients who have already been prescribed at least one opioid. Both of these choices artificially bypass the main challenge of imbalance. Thus, our study is more in line with Ellis et al. (2019), who attempt to deal with the original imbalanced dataset of the entire patient population. Ellis et al. (2019) report an AUC of 0.82 for an outcome similar to the *Overdose* outcome, for which we achieve an AUC of 0.87. As mentioned earlier, there are very few studies building prediction models for *FatalOverdose*. Ferris et al. (2019) reports an AUC of 0.81 for this task, but their study population considers only patients who have received at least one opioid prescription. Using data on all-payer hospital discharges, the prescription drug monitoring program, public-sector specialty behavioral treatment, and criminal justice records for property or drug-associated offenses, Saloner et al. (2020) build a predictive model for *FatalOverdose* with an AUC of 0.89 over a one-year prediction window. When using only hospital and prescription drug monitoring data, as we do, Saloner et al. (2020) report an AUC of 0.86 for a one-year prediction window. In §3.2.2, we consider a one-year prediction window and find that the AUC of our model for *FatalOverdose* is 0.88, therefore providing a modest improvement over Saloner et al. (2020) when a fair comparison is attempted.

Table 2.4 summarizes these comparisons along with some additional benchmarks. Based on this comparison, we conclude that our parsimonious modeling approach improves upon the state of the art models for predicting opioid-related adverse events.

Using a parsimonious approach also has other important advantages. Data from medical insurance claims or EHRs typically do not contain information to identify the

Table 2.4: AUC comparison with results in prior literature

Outcome	Our AUC	Comparable work	Best AUC reported
AnyOpioidHarm	0.95	Hasan et al. (2019)	0.97
		Hastings et al. (2020)	0.80, but for patients with at least one opioid prescription
Overdose	0.87	Dong et al. (2019)	0.95, but in a dataset with undersampled negative obs.
		Ellis et al. (2019)	0.82
		Geissert et al. (2018)	0.80, but for patients with at least one opioid prescription
		Lo-Ciganic et al. (2019)	0.89, but for patients with at least one opioid prescription
FatalOverdose	0.83	Ferris et al. (2019)	0.81, but for patients with at least one opioid prescription
		Saloner et al. (2020)	0.89, but for a one-year prediction window

outcome of *FatalOverdose*. (We obtained information on the *FatalOverdose* outcome using a dataset from the District Attorney's office.) We believe this is a key reason for the scarcity of existing works on predicting *FatalOverdose*. Further, even when this data is available (such as in Geissert et al. (2018)), the minimal number of positive observations for the *FatalOverdose* outcome makes it challenging to build good predictive models. Our parsimonious model illustrates that even in the absence of data on *FatalOverdose* or with a highly imbalanced dataset, it is possible to train models and operationalize programs aimed at dealing with *FatalOverdose*.

Our modeling approach is also in line with various other efforts to build models for opioid risk prediction that seek parsimony. Geissert et al. (2018) and Hasan et al. (2019) consider finding a small set of features that can predict opioid-related outcomes. In addition to building models that use all claims data, Dong et al. (2019) consider building models without using diagnosis codes from claims. As we are interested in predicting multiple opioid-related outcomes, our search for a parsimonious approach focuses on training a single model that can be used for planning interventions for all three outcomes. This decrease in complexity from maintaining only one model is beneficial from an implementation perspective.

Chapter 3

Operationalizing risk stratification models

The previous chapter follows the literature on opioid-related risk stratification models in aiming to maximize AUC, a standard measure of predictive model performance. However, from the perspective of operationalizing such models in a healthcare system, this analysis has two key limitations: it does not quantify the resources needed for planning effective interventions, and it ignores important implementation issues, such as a lag in obtaining data, or a need for different models for different kind of interventions. To facilitate implementation and assist managerial decision making, we further build on our analysis to address these two limitations in §3.1 and §3.2 respectively.

3.1 Capacity planning

Effective and timely interventions on the "right" patients have the potential for preventing future opioid-related adverse outcomes. In practice, the model developed in §2.4 can be used by a healthcare organization at regular time intervals to estimate the risk of an adverse outcome for each patient, stratify patients by risk level, and then intervene with the highest-risk patients. The objective of any intervention is to intervene with a patient prior to an adverse opioid-related event occurring, which we refer to as *catching* the adverse event. For an organization with limited resources, this raises the question of how large the high-risk cohort needs to be to catch a targeted fraction of outcomes, because a larger cohort implies a larger required number of social workers, counselors, and psychologists. In this section, we present a counterfactual analysis, based on our data, to generate bounds on the capacity requirements for such interventions.

The number of opioid-related outcomes prevented by an intervention program can be viewed as being impacted by two key factors: appropriate identification of high-risk patients, and effectiveness of the intervention in preventing future opioid-related adverse outcomes for a given patient. The focus of our analysis is exclusively on the impact of the first factor. To abstract away the impact of the intervention effectiveness, our counterfactual analysis assumes that the intervention in question is fully effective only for a time window of certain length. By varying the length of that time window, our approach enables us to derive approximate bounds on the desired capacity requirements for intervention programs.

In our counterfactual analysis, we assume that our risk stratification model is reapplied to the patient population every 90 days, to update the categorization of patients as high- or low-risk for opioid-related harm. A high-risk cohort is formed at the beginning of each period by ranking all patients in order of their predicted risk of an opioid-related outcome and then taking the subset of patients at highest risk, such that the size of the cohort is a prespecified fraction of all patients. The provider organization uses its resources to intervene with the high-risk cohort during the following 90-day interval. Therefore the fraction of the population defining the high-risk cohort corresponds to the capacity requirement.

To develop bounds on the capacity required for the intervention program, we run two sets of analyses. First, we assume that the effect of the intervention is short-lasting and only catches adverse events occurring during exactly the 90 days that a patient is receiving the intervention. This provides an upper bound on the required intervention capacity since patients would only be protected from adverse events while they are actively receiving the intervention (§3.1.1). Second, we assume that the intervention has a long-lasting effect, so that patients who have received the intervention at any point in time in the past are protected from adverse events. This lasting intervention would therefore catch all future adverse events, once it has been provided to a patient. This provides a lower bound on the required intervention capacity since each patient only needs to receive the intervention at most once (§3.1.2). For each analysis, we examine what proportion of patients need to receive the intervention in order for a target fraction of adverse events to be caught.

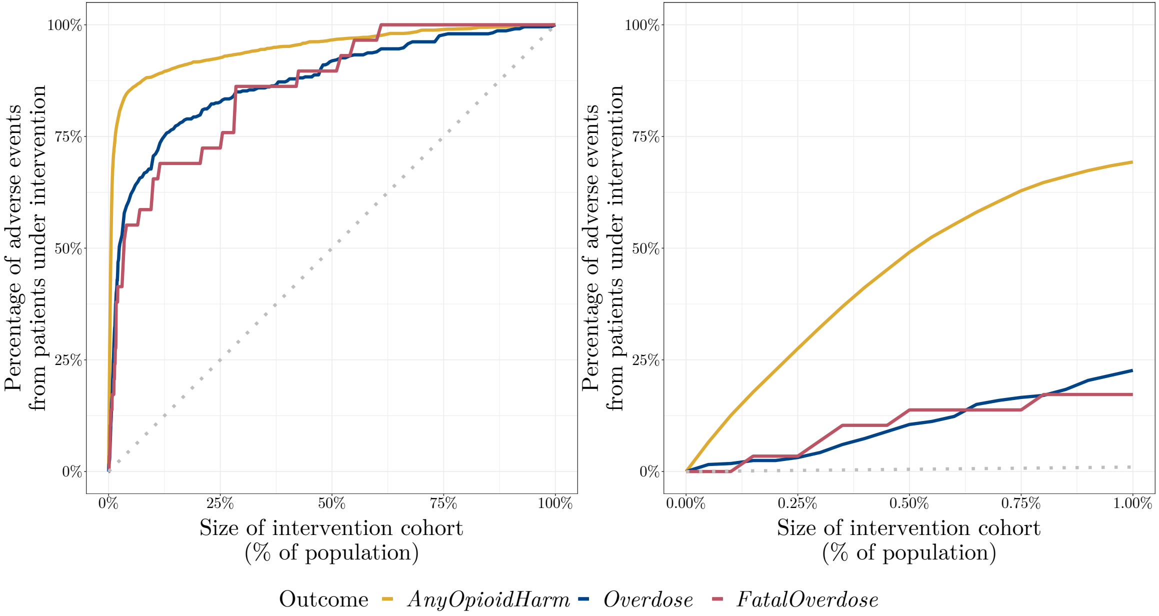
3.1.1 Upper bound on capacity requirements

Suppose that every 90 days when the risk stratification model is reapplied to the patient population, some proportion of the highest risk patients receive an intervention that protects them from adverse opioid-related events for exactly the period of 90 days until the next risk stratification.

To shed light on the capacity requirements of such an intervention so that it can catch a targeted percentage of adverse outcomes, we backtest our models using historical information in the testing set. Specifically, we begin each 90-day period stratifying patients based on information available at the beginning of the period and then form the high-risk cohort using a certain fraction of the patient population. We then focus on all outcomes that occurred during this period, and record the fraction of these incidents (either *AnyOpioidHarm*, or *Overdose*, or *FatalOverdose*) that were associated with patients from the high-risk cohort—in which case, we consider the incident caught by the intervention.

In Figure 3-1, we plot the fraction of incidents caught by the intervention program as we vary the fraction of patients who form the high-risk cohort. For both *AnyOpioidHarm* and *Overdose*, we observe that there is a steep rise in the curve close to 0. With an intervention cohort of size only 1% of the patient population, more than 17% of *FatalOverdose*, 22% of *Overdose* and 69% of *AnyOpioidHarm* incidents are caught. When the size of this cohort is increased to 15%, the number of incidents caught increases to more than 65%, 75% and 90%, respectively. Thus, for all outcomes, we observe that even by providing a short-lasting intervention to a relatively small

Figure 3-1: Fraction of adverse opioid-related outcomes caught in the intervention cohort, under an intervention with 90-day duration



Notes: Plot on the right is a zoomed-in version of the plot on the left. Performance of an intervention scheme based on a random classifier is presented as a dotted gray line.

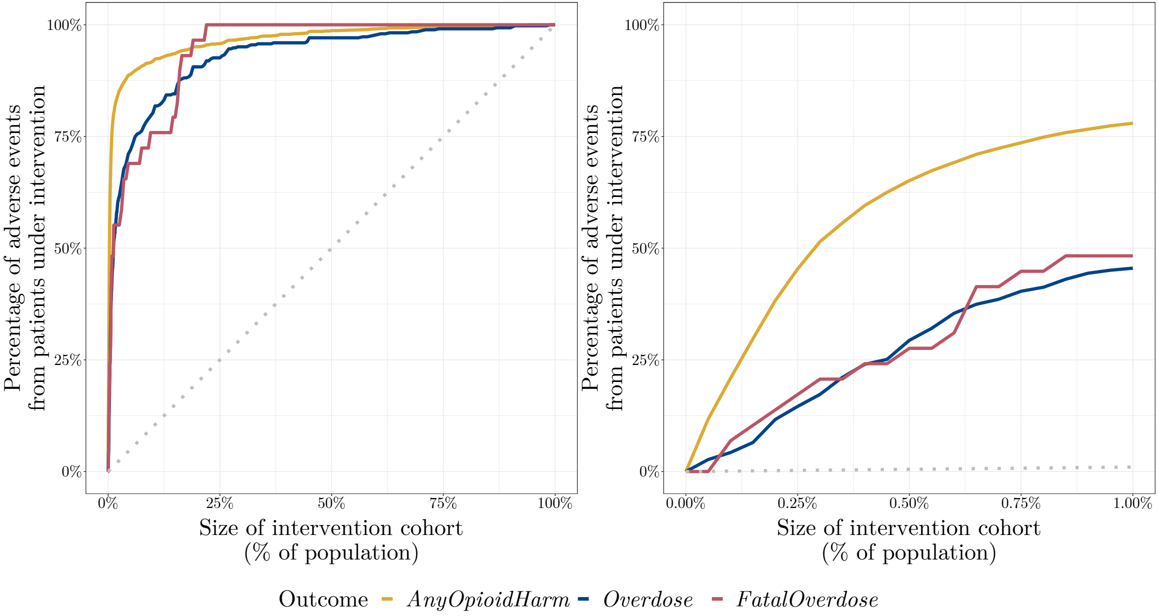
fraction of the population, the intervention can catch a high number of opioid-related adverse events. Intuitively, since prediction is more challenging for the *FatalOverdose* outcome, a larger cohort needs to receive the intervention to catch such instances.

3.1.2 Lower bound on capacity requirements

Now suppose that every 90 days, when the risk stratification model is reapplied and a new set of the highest risk patients receives an intervention, this intervention has a lasting effect that protects patients from any future opioid-related adverse events. Under this scenario, what proportion of patients should be used to construct the high-risk cohort, in order for the intervention to catch a targeted fraction of patients who will experience opioid-related adverse events?

To conduct our analysis, we use a similar backtesting strategy as in the previous section. Specifically, in each period we begin by stratifying patients based on information available at the beginning of the period and then form the high-risk cohort using a certain fraction of the patient population. Patients in the high-risk cohort are

Figure 3-2: Fraction of adverse opioid-related outcomes caught in the intervention cohort, under an intervention that prevents any future adverse outcomes



Notes: Plot on the right is a zoomed-in version of the plot on the left. Performance of an intervention based on a random classifier is presented as a dotted gray line.

intervened upon and are subsequently considered to be under the effect of the intervention for the rest of the horizon. We consider all opioid-related events that occur for patients when they are under the effect of the intervention to be caught. We then record the fraction of all adverse events that were caught due to the intervention.

In Figure 3-2, we plot this fraction as a function of the percentage of patients intervened upon. Similar to the case of the short-lasting intervention, we observe that intervening on a small fraction of the patients can prevent a large fraction of opioid outcomes from occurring, especially for *AnyOpioidHarm*. This is even clearer from the plot on the right in Figure 3-2, which zooms in to the region where the percentage of patients intervened on is less than 1%. We observe that just by intervening on 1% of the patients, more than 77%, 45% and 48% of *AnyOpioidHarm*, *Overdose* and *FatalOverdose* outcomes can be averted, respectively.

Table 3.1: Bounds on the intervention capacity required to prevent a targeted fraction of adverse outcomes

	(1)	(2)	(3)
Target	*AnyOpioidHarm*	*Overdose*	*FatalOverdose*
10%	0.1% - 0.1%	0.2% - 0.5%	0.2% - 0.4%
20%	0.1% - 0.2%	0.4% - 0.9%	0.3% - 1.2%
30%	0.2% - 0.3%	0.5% - 1.3%	0.6% - 1.7%
40%	0.2% - 0.4%	0.8% - 1.7%	0.7% - 2.1%
50%	0.3% - 0.5%	1.2% - 2.4%	0.9% - 3.5%
60%	0.4% - 0.7%	2.2% - 4.5%	3.0% - 7.0%
70%	0.6% - 1.1%	4.5% - 10.0%	4.5% - 11.5%
80%	1.2% - 2.4%	10.0% - 20.5%	14.5% - 25.5%
90%	6.5% - 14.0%	19.0% - 47.5%	16.0% - 42.5%

Notes: Columns (1), (2), and (3) provide lower and upper bounds for the proportion of the population that needs to receive an intervention to catch a given target fraction of adverse events. Lower and upper bounds are based on the data underlying the curves in Figures 3-1 and 3-2, respectively.

3.1.3 Capacity planning bounds

Taken together, the results from the two previous subsections provide upper and lower bounds on the intervention capacity required to catch a given fraction of adverse events. Specifically, §3.1.1 provides an upper bound on the required intervention capacity since it assumes we only catch adverse events for exactly the 90 days when a patient is receiving an intervention. Similarly, §3.1.2 provides a lower bound on the same capacity, as it assumes that an intervention can catch all future adverse events for a patient by intervening once.

Table 3.1 summarizes the insights from this interpretation. It demonstrates that in order to catch 50% of instances of *AnyOpioidHarm*, for example, somewhere between 0.3% (if the intervention is lasting) and 0.5% (if the intervention is only effective for one period of 90 days) of the population would need to be enrolled in a prevention program. The analogous statistics for *Overdose* and *FatalOverdose* are 1.2% - 2.4% and 0.9% - 3.5%, respectively. This demonstrates the impact that risk stratification can have on the cost-benefit trade-off for an intervention program. By enrolling a very small proportion of the population in prevention programs, a large proportion of adverse events can be caught.

3.2 Implementation tradeoffs

In this section, we consider three different decisions that a healthcare organization may have to make in the deployment of opioid risk predictive models, where each decision affects multiple conflicting objectives. In particular, each decision involves a tradeoff between predictive performance and another desirable objective, and we quantify the change in predictive performance that may result. So doing could assist healthcare organizations, like SI PPS, to navigate the underlying tradeoffs more effectively. The decisions that we study concern (1) whether to use an interpretable model, which can aid both in intervention design and face validity; (2) whether to use a model with a shorter or longer prediction window; and (3) whether to invest in IT infrastructure that reduces the lag in available data.

3.2.1 Interpretability

Healthcare practitioners often prefer predictive models that are interpretable (Ahmad et al., 2018). These models provide a more intuitive understanding of the prediction process, which can enable practitioners to validate the model and aid in choosing appropriate interventions. Further, interpretability can help in making sure that the model is not biased and does not derive its predictive power from sensitive features of the patient. Predictive models that improve interpretability, however, sometimes come at the cost of compromised predictive performance. Here we shed light on this tradeoff in the context of our study by quantifying how much predictive performance is lost when we fit an interpretable model rather than an XGBoost model.

We train an interpretable model for *AnyOpioidHarm* using an OCT (Bertsimas and Dunn, 2017; Interpretable AI, LLC, 2020). An OCT is a single decision tree, which provides a transparent view of its working. Whereas standard techniques for learning decision trees such as CART (Breiman et al., 1984) build the decision tree in

Table 3.2: AUC of OCT for different outcomes

Outcome	AUC
AnyOpioidHarm	0.93 (0.93 - 0.94)
Overdose	0.84 (0.81 - 0.86)
FatalOverdose	0.74 (0.62 - 0.86)

Notes: Parentheses report 95% confidence intervals.

sequential greedy steps, all the node splits in an OCT are simultaneously optimized. This optimization is possible because of recent advancements in solvers for mixed integer optimization, and can produce trees with higher accuracy (Bertsimas and Dunn, 2017). The high accuracy and interpretability make OCT attractive for medical applications.

The AUCs of this model for all three outcomes are presented in Table 3.2. We observe that the performance is similar to the parsimonious XGBoost model, although there is slight deterioration in AUC across outcomes. Thus, this analysis suggests that with little compromise on accuracy, these models can be effectively used when there is a preference for interpretability. Appendix C contains details on the training of the OCT model as well as a depiction of the resulting decision tree (Figure C-1).

3.2.2 Prediction window length

In the prediction of opioid-related adverse outcomes, different prediction windows can be of interest depending on the type of intervention under consideration, and the outcome which we would like to prevent. Some interventions aim for long-term behavioral changes whereas others can aim for preventing immediate adverse outcomes, particularly for severe harm such as *FatalOverdose*. The responsiveness required in the latter scenario is best supported by a model with a shorter prediction window, but the tradeoff is that we expect prediction of an adverse outcome to be more difficult over a shorter period of time. We quantify this tradeoff by evaluating the performance of parsimonious XGBoost models trained for different prediction windows: 15 days, 30 days, 90 days, and 365 days.

Table 3.3: AUC of XGBoost models with different prediction windows

	Prediction window			
	15 days	30 days	90 days	365 days
AnyOpioidHarm	0.98 (0.98 - 0.98)	0.97 (0.97 - 0.98)	0.95 (0.95 - 0.95)	0.94 (0.94 - 0.95)
Overdose	0.90 (0.87 - 0.94)	0.88 (0.85 - 0.91)	0.87 (0.85 - 0.89)	0.91 (0.89 - 0.92)
FatalOverdose	0.82 (0.60 - 1.00)	0.81 (0.69 - 0.94)	0.83 (0.73 - 0.92)	0.88 (0.81 - 0.94)

Notes: Parentheses report 95% confidence intervals.

To characterize the performance of these models, we report the AUC with varying prediction windows in Table 3.3. We observe that fairly high AUC values (all above 0.94) are attained by the models predicting *AnyOpioidHarm*. This implies that patients at a high risk of having *AnyOpioidHarm* can be identified quite accurately both over shorter and longer prediction windows. For both *Overdose* and *FatalOverdose* outcomes, making predictions over the longest prediction window has the highest AUC. However, the AUC values for both outcomes are still fairly high across all prediction windows (all above 0.87 and 0.81, respectively). We conclude that the parsimonious model loses relatively little predictive performance when varying the prediction window.

3.2.3 Data delay

In practice, the prediction model would be used periodically to stratify patients by risk level, but the most recent data might not be available for constructing predictor variables. Due to a lag in claims submission by providers, most databases deem data to be usable only after 90 days (Majumder and Rose, 2020). A lag in the dataset becomes even more likely when the model dependss on data from multiple sources, such as EHRs, insurance claims, and emergency medical services. A delay in any one of the sources leads to a lag in the data available when using the prediction models. This in turn could compromise predictive power. A healthcare organization may consider investing money in their IT infrastructure to reduce this delay, but how much would predictive performance be improved by this investment?

To capture the tradeoff between predictive performance and delay in data avail-

Table 3.4: AUC of XGBoost models having delay in access to data

	Data Delay			
	0 days	30 days	90 days	180 days
AnyOpioidHarm	0.95 (0.95 - 0.95)	0.95 (0.95 - 0.96)	0.94 (0.94 - 0.94)	0.94 (0.93 - 0.94)
Overdose	0.87 (0.85 - 0.89)	0.89 (0.88 - 0.91)	0.87 (0.85 - 0.89)	0.87 (0.85 - 0.89)
FatalOverdose	0.83 (0.73 - 0.92)	0.84 (0.77 - 0.92)	0.83 (0.73 - 0.92)	0.86 (0.77 - 0.94)

Notes: Parentheses report 95% confidence intervals.

ability, we construct datasets incorporating different lags (30, 60, and 180 days) in accessing the data. In these datasets, each predictor variable is calculated using the information up to 30, 60, or 180 days before the observation date. For example, if there is a 30-day delay in accessing data, an observation on the date May 1, 2020 has predictor variables that only use the patient history prior to April 1, 2020. We then train and test a parsimonious XGBoost model on each of these datasets.

The performance of models trained on these datasets with different data delays is presented in Table 3.4. For *AnyOpioidHarm*, the AUC drops from 0.95 when there is no delay in the data to 0.94 when there is a 180-day delay. For all the three outcomes, we observe that as the data delay increases up to 180 days, the change in predictive performance is modest. Thus, the prediction models have the desirable property of being robust to delays in the data, suggesting that investments in IT infrastructure to reduce data delay may not be worthwhile.

Chapter 4

Conclusion

The opioid epidemic caused more than 100 deaths per day in the US in 2018 (Wilson et al., 2020). In addition to the impact on human life, the societal costs of OUD and fatal opioid overdoses were estimated to be $1.02 trillion in the US in 2017, including the costs of healthcare, criminal justice, and lost productivity (Florence et al., 2021). To counter this problem, various kinds of policy measures have been taken (National Academies of Sciences, Engineering, and Medicine, 2017). This includes personalized preventive interventions aimed at individuals showing early symptoms or who are predicted to be at a high level of risk, referred to as *indicated interventions* (National Research Council and Institute of Medicine, 2009). These methods tailor prevention and treatment by targeting them to the right patient at the right time. In planning these interventions, a key step is to identify high-risk patients. In this work, we present machine learning models to predict patient-level risk of various opioid-related adverse outcomes. Then, we focus on various aspects of operationalization of these models, including capacity planning for intervention, and tradeoffs that arise from implementation considerations.

We show that machine learning models can be highly accurate for predicting the risk of opioid-related adverse outcomes. When fitting an XGBoost model to each of the outcomes *AnyOpioidHarm*, *Overdose*, and *FatalOverdose*, the least severe outcome (*AnyOpioidHarm*) can be predicted with the highest accuracy, whereas it is most challenging to predict the most severe outcome (*FatalOverdose*). We propose an alternative parsimonious modeling approach in which the model trained to predict *AnyOpioidHarm* is also used to predict *Overdose* and *FatalOverdose* and find that this improves AUC, particularly for prediction of *FatalOverdose.* This not only eases implementation, since a healthcare organization would have to maintain only one model rather than three, it also means that an organization can predict *FatalOverdose* even without a data source recording deaths. Indeed, fatal opioid overdoses often are not available in databases of health insurance claims or EHRs which are most commonly used for building these risk prediction models.

We quantify the value of these prediction models in identifying a cohort with whom to conduct interventions. We show that by forming a cohort using a small fraction of patients who are estimated to be at high risk based on the above models, a large fraction of future opioid-related outcomes would be potentially averted. Using these prediction models as the basis for preventive interventions is therefore attractive even in resource-constrained scenarios.

These machine learning models have multiple further desirable properties, in that their predictive performance is robust to tradeoffs with other desirable objectives. With very little compromise on accuracy, an easily interpretable decision tree model can also be used to obtain risk estimates. As interventions might vary in the duration they are effective for, we build separate models for varying prediction windows, and find little change in predictive performance. Model performance is also robust to delays in the availability of data.

Our dataset imposes some potential limitations on the results of our analysis. Both our predictor variables and outcomes variables may be potentially censored. We also do not have data from clinical notes which could provide additional information beyond the structured data of diagnoses, prescriptions, encounters, and demographics. Although these limitations may limit predictive power, they do not bias our analysis because the available data mimics the data environment in which the models would actually be deployed.

Our work opens up multiple avenues for further investigation on the role of indicated interventions in controlling the opioid epidemic. To more accurately estimate the impact of an intervention program based on a predictive model, we would need causal insights on how the intervention changes opioid disease progression, and how long the intervention is effective for. These insights can be used to build a detailed model for an appropriate timeline of conducting the intervention. Predictive models can also be developed to predict which intervention will be effective on which patient, further helping personalize interventions.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix A

Codes and variables for dataset construction

The tables in this appendix provide further detail on the codes and definitions that we used to transform our raw data into a dataset for analysis, as described in §2.2.

Table A.1: ICD codes for cancer

Code	Description
ICD-9:	
140–149	Malignant neoplasms of lip, oral cavity, and pharynx
150–159	Malignant neoplasms of digestive organs and peritoneum
160–165	Malignant neoplasms of respiratory and intrathoracic organs
170–175	Malignant neoplasms of bone, connective tissue, skin, and breast
176–176	Kaposi's sarcoma
179–189	Malignant neoplasms of genitourinary organs
190–199	Malignant neoplasms of other and unspecified sites
200–208	Malignant neoplasms of lymphatic and hematopoietic tissue
ICD-10:	
C00–C14	Malignant neoplasms of lip, oral cavity, and pharynx
C15–C26	Malignant neoplasms of digestive organs
C30–C39	Malignant neoplasms of respiratory and intrathoracic organs
C40–C41	Malignant neoplasms of bone and articular cartilage
C43	Malignant melanoma of skin
C4A	Merkel cell carcinoma
C44	Other malignant neoplasms of skin
C45–C49	Malignant neoplasms of mesothelial and soft tissue
C50	Malignant neoplasms of breast
C51–C68	Malignant neoplasms of genitourinary organs
C69–C72	Malignant neoplasms of eye, brain, and other parts of central nervous system
C73–C75	Malignant neoplasms of thyroid and other endocrine glands
C7A-C7B	Neuroendocrine tumors
C76–C80	Malignant neoplasms of ill-defined, other secondary and unspecified sites
C81–C96	Malignant neoplasms of lymphoid, hematopoietic and related tissue
D37–D48	Neoplasms of uncertain behavior, polycythemia vera and myelodysplastic syndromes
D49	Neoplasms of unspecified behavior
Q85.0	Neurofibromatosis (nonmalignant)

Notes: ICD-9 codes are from Lo-Ciganic et al. (2019), ICD-10 codes are from Office of Inspector General (2018).

Table A.2: Predictor variables and their definitions

Variable	Description
Previous_Harm_N	No. of previous incidents of *AnyOpioidHarm*
Previous_PDAMAT_N	No. of previous incidents of *Overdose*, opioid dependence, opioid abuse, or filled prescriptions for MAT
Previous_PD_N	No. of previous incidents of *Overdose* or opioid dependence
Previous_Overdose_N	No. of previous incidents of *Overdose*
Gender	Female, male, unknown, or missing
Race	Asian, Black, Hispanic, Native American, White, other, or missing
Ethnicity	Hispanic or Latino; not Hispanic or Latino; unknown; or missing
Age	Age in years
Outpatient_N	No. of outpatient encounters in past 90 days
ED_N	No. of emergency department encounters in past 90 days
Inpatient_N	No. of inpatient encounters in past 90 days
Schedule_2_Ind	Indicator of any schedule 2 prescription fills in past 90 days
Schedule_3_Ind	Indicator of any schedule 3 prescription fills in past 90 days
Schedule_4_Ind	Indicator of any schedule 4 prescription fills in past 90 days
Schedule_5_Ind	Indicator of any schedule 5 prescription fills in past 90 days
Opioid_Benzo_Days	Total days of co-prescription of opioids and benzodiazepines in past 90 days
	No. of prescription fills in past 90 days, of
Antidepressant_N	Antidepressants
Benzo_N	Benzodiazepines
Gabapentin_N	Gapapentin
Muscle_Relaxant_N	Muscle relaxants
Pregabalin_N	Pregabalin
Buprenorphine_OUD_N	Buprenorphine variants which are specifically used to treat OUD
Methadone_N	Methadone
Naltrexone_N	Naltrexone
Buprenorphine_LA_N	Buprenorphine (long-acting)
Butorphanol_SA_N	Butorphanol (short-acting)
Codeine_SA_N	Codeine (short-acting)
Dihydrocodeine_SA_N	Dihydrocodeine (short-acting)
Fentanyl_LA_N	Fentanyl (long-acting)
Fentanyl_SA_N	Fentanyl (short-acting)
Hydrocodone_LA_N	Hydrocodone (long-acting)
Hydrocodone_SA_N	Hydrocodone (short-acting)
Hydromorphone_LA_N	Hydromorphone (long-acting)
Hydromorphone_SA_N	Hydromorphone (short-acting)
Levomethadyl_LA_N	Levomethadyl (long-acting)
Levorphanol_LA_N	Levorphanol (long-acting)
Meperidine_SA_N	Meperidine (short-acting)
Morphine_LA_N	Morphine (long-acting)
Morphine_SA_N	Morphine (short-acting)
Opium_SA_N	Opium (short-acting)
Oxycodone_LA_N	Oxycodone (long-acting)

Table A.3: Predictor variables and their definitions (continued)

Variable	Description
	No. of prescription fills in past 90 days, of
Oxycodone_SA_N	Oxycodone (short-acting)
Oxymorphone_LA_N	Oxymorphone (long-acting)
Oxymorphone_SA_N	Oxymorphone (short-acting)
Pentazocine_SA_N	Pentazocine (long-acting)
Propoxyphene_SA_N	Propoxyphene (long-acting)
Tapentadol_LA_N	Tapentadol (long-acting)
Tapentadol_SA_N	Tapentadol (short-acting)
Tramadol_LA_N	Tramadol (long-acting)
Tramadol_SA_N	Tramadol (short-acting)
	Total days supply of prescription fills in past 90 days, of
Antidepressant_DS	Antidepressants
Benzo_DS	Benzodiazepines
Gabapentin_DS	Gabapentin
Muscle_Relaxant_DS	Muscle relaxants
Pregabalin_DS	Pregabalin
Buprenorphine_OUD_DS	Buprenorphine variants which are specifically used to treat OUD
Methadone_DS	Methadone
Naltrexone_DS	Naltrexone
Buprenorphine_LA_DS	Buprenorphine (long-acting)
Butorphanol_SA_DS	Butorphanol (short-acting)
Codeine_SA_DS	Codeine (short-acting)
Dihydrocodeine_SA_DS	Dihydrocodeine (short-acting)
Fentanyl_LA_DS	Fentanyl (long-acting)
Fentanyl_SA_DS	Fentanyl (short-acting)
Hydrocodone_LA_DS	Hydrocodone (long-acting)
Hydrocodone_SA_DS	Hydrocodone (short-acting)
Hydromorphone_LA_DS	Hydromorphone (long-acting)
Hydromorphone_SA_DS	Hydromorphone (short-acting)
Levomethadyl_LA_DS	Levomethadyl (long-acting)
Levorphanol_LA_DS	Levorphanol (long-acting)
Meperidine_SA_DS	Meperidine (short-acting)
Morphine_LA_DS	Morphine (long-acting)
Morphine_SA_DS	Morphine (short-acting)
Opium_SA_DS	Opium (short-acting)
Oxycodone_LA_DS	Oxycodone (long-acting)
Oxycodone_SA_DS	Oxycodone (short-acting)
Oxymorphone_LA_DS	Oxymorphone (long-acting)
Oxymorphone_SA_DS	Oxymorphone (short-acting)
Pentazocine_SA_DS	Pentazocine (short-acting)
Propoxyphene_SA_DS	Propoxyphene (short-acting)
Tapentadol_LA_DS	Tapentadol (long-acting)
Tapentadol_SA_DS	Tapentadol (short-acting)
Tramadol_LA_DS	Tramadol (long-acting)
Tramadol_SA_DS	Tramadol (short-acting)

Table A.4: Predictor variables and their definitions (continued)

Variable	Description
	No. of diagnostic codes entered in past year, for
Opioid_Abuse_N	Opioid abuse
Opioid_Dependence_N	Opioid dependence
Opioid_Use_N	Unspecified use of opioids
Opioid_Adverse_Effects_N	Adverse effects due to opioids
Opioid_Poisoning_N	Opioid poisoning
Drug_Dependence_N	Dependence on drug(s) other than opioids
Drug_Abuse_N	Abuse of drug(s) other than opioids
Drug_Other_Problems_N	Other problems involving drug(s) other than opioids
Drug_Use_N	Unspecified use of drug(s) other than opioids
Drug_Adverse_Effects_N	Adverse effects due to drug(s) other than opioids
Drug_Poisoning_N	Poisoning by drug(s) other than opioids
Alcohol_Dependence_N	Alcohol dependence
Alcohol_Abuse_N	Alcohol abuse
Alcohol_Use_N	Unspecified use of alcohol
Alcohol_Other_Problems_N	Other alcohol-related problems
Alcohol_Poisoning_N	Alcohol poisoning
Anxiety_N	Anxiety
Bipolar_Disorder_N	Bipolar disorder
Personality_Disorders_N	Personality disorders
Depression_N	Depression
ADHD_N	Attention deficit hyperactivity disorder (ADHD)
Schizophrenia_N	Schizophrenia
Psychotic_Disorders_N	Psychotic disorders other than schizophrenia

Table A.5: NDC codes for non-opioid drugs

Drug class	Source of codes
Antidepressants	National Committee for Quality Assurance (2018)
Benzodiazepines	National Center for Injury Prevention and Control (2018)
Muscle relaxants	National Center for Injury Prevention and Control (2018)
Gabapentinoids	American Society of Health-System Pharmacists, Inc. (2019)
Pregabalin	American Society of Health-System Pharmacists, Inc. (2019)
Naltrexone	Chronic Condition Data Warehouse (2019)

Notes: *AHFS ® Pharmacologic/Therapeutic Classification* © used with permission. ©2020, the American Society of Health-System Pharmacists, Inc. (ASHP). The Data is a part of the *AHFS Drug Information* ®; ASHP is not responsible for the accuracy of transpositions from the original context.

Table A.6: ICD codes for opioid-related harm

Type of harm	Code	Description
Unspecified use	F11.9	Opioid use, unspecified
Opioid abuse	305.5[012]	Opioid abuse
	F11.1[024589]	Opioid abuse
Opioid dependence	304.0[012]	Opioid type dependence
	304.7[012]	Combinations of opioid type drug with any other drug dependence
	F11.2[024589]	Opioid dependence
Adverse effects of opioids	E935.0	Heroin causing adverse effects in therapeutic use
	E935.1	Methadone causing averse effects in therapeutic use
	E935.2	Other opiates and related narcotics causing adverse effects in therapeutic use
	E940.1	Opiate antagonists causing adverse effects in therapeutic use
	T40.0X5	Adverse effect of opium
	T40.2X5	Adverse effect of other opioids
	T40.3X5	Adverse effect of methadone
	T40.4X5	Adverse effect of other synthetic narcotics
	T40.605	Adverse effect of unspecified narcotics
	T40.695	Adverse effect of other narcotics
Opioid poisoning	970.1	Poisoning by opiate antagonists
	965.00	Poisoning by opium (alkaloids), unspecified
	965.01	Poisoning by heroin
	965.02	Poisoning by methadone
	965.09	Poisoning by other opiates and related narcotics
	E850.0	Accidental poisoning by heroin
	E850.1	Accidental poisoning by methadone
	E850.2	Accidental poisoning by other opiates and related narcotics
	T40.0X[14]	Poisoning by opium
	T40.1X[14]	Poisoning by heroin
	T40.2X[14]	Poisoning by other opioids
	T40.3X[14]	Poisoning by methadone
	T40.4X[14]	Poisoning by other synthetic narcotics
	T40.60[14]	Poisoning by unspecified narcotics
	T40.69[14]	Poisoning by other narcotics

Notes: Codes that begin with a digit or the letter E are ICD-9 codes, and the remaining codes are ICD-10 codes. Square brackets indicate that any one of the digits within the brackets may occupy that position in the code. Codes are derived from Heslin et al. (2017).

Appendix B

XGBoost classification model

In §2.5, we reported the performance of XGBoost models trained using a ranking objective. We additionally train an XGBoost model with a classification objective, using the parismonious approach (i.e., a single model trained on the *AnyOpioidHarm* outcome). We report its performance in Table B.1, and find that it is similar to the results obtained for the ranking-based parsimonious model (Table 2.3).

Table B.1: AUC of XGBoost classification model

Outcome	AU-ROC
AnyOpioidHarm	0.95 (0.95 - 0.95)
Overdose	0.87 (0.86 - 0.89)
FatalOverdose	0.81 (0.69 - 0.92)

Notes: Parentheses report 95% confidence intervals.

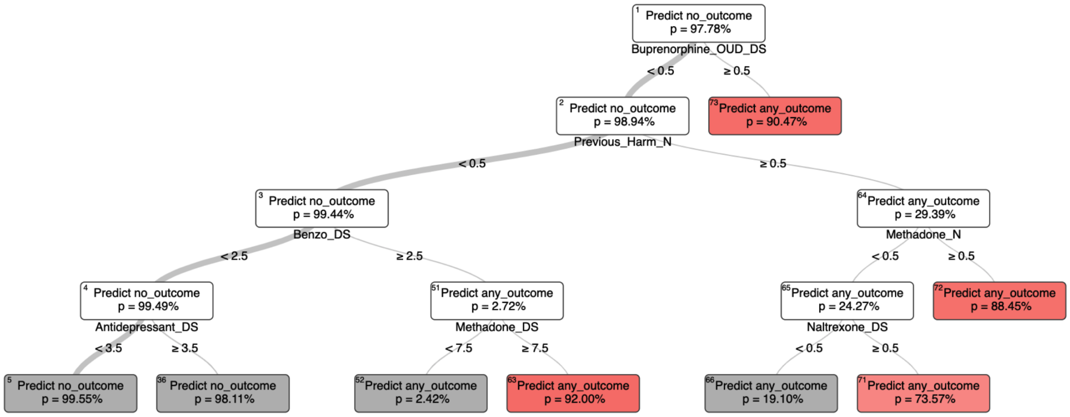
THIS PAGE INTENTIONALLY LEFT BLANK

Appendix C

OCT training

The OCT model in §3.2.1 takes a much longer time to train as compared to the XGBoost models. Thus, we sub-sample the training data for use by the OCT model. In particular, we randomly sample half of the observations from the full training dataset and use the 50 most important features as reported by the XGBoost model to construct the training dataset for the OCT model. The resulting decision tree itself is pictured in Figure C-1. We stress that the variables selected for the tree are predictive, but do not imply a causal effect on outcomes.

Figure C-1: Optimal classification tree


1 Predict no_outcome p = 97.78%
Buprenorphine_OUD_DS
< 0.5
≥ 0.5
2 Predict no_outcome p = 98.94%
Previous_Harm_N
73 Predict any_outcome p = 90.47%
< 0.5
≥ 0.5
3 Predict no_outcome p = 99.44%
Benzo_DS
64 Predict any_outcome p = 29.39%
Methadone_N
< 2.5
≥ 2.5
< 0.5
≥ 0.5
4 Predict no_outcome p = 99.49%
Antidepressant_DS
51 Predict any_outcome p = 2.72%
Methadone_DS
65 Predict any_outcome p = 24.27%
Naltrexone_DS
72 Predict any_outcome p = 88.45%
< 3.5
≥ 3.5
< 7.5
≥ 7.5
< 0.5
≥ 0.5
5 Predict no_outcome p = 99.55%
36 Predict no_outcome p = 98.11%
52 Predict any_outcome p = 2.42%
63 Predict any_outcome p = 92.00%
66 Predict any_outcome p = 19.10%
71 Predict any_outcome p = 73.57%

Notes: This is a visualization of the first few layers of the optimal classification tree. *no_outcome* stands for predicting no occurrence of *AnyOpioidHarm* and *any_outcome* stands for predicting an occurrence of *AnyOpioidHarm*. The nodes shaded gray denote a collapsed subset of the tree whereas the nodes shaded red are leaf nodes.

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 559–560, 2018.

American Society of Health-System Pharmacists, Inc. AHFS Pharmacologic-Therapeutic Classification System, 2019.

Charles E. Argoff and Daniel I. Silvershein. A comparison of long- and short-acting opioids for the treatment of chronic noncancer pain: Tailoring therapy to meet patient needs. *Mayo Clinic Proceedings*, 84(7):602–612, 2009.

Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, 2017.

Rok Blagus and Lara Lusa. Evaluation of SMOTE for high-dimensional class-imbalanced microarray data. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 89–94. IEEE, 2012.

Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. *Classification and Regression Trees.* CRC Press, 1984.

Stephen F. Butler, Simon H. Budman, Kathrine Fernandez, and Robert N. Jamison. Validation of a screener and opioid assessment measure for patients with chronic pain. *Pain*, 112(1-2):65–75, 2004.

Hsien Yen Chang, Noa Krawczyk, Kristin E. Schneider, Lindsey Ferris, Matthew Eisenberg, Tom M. Richards, B. Casey Lyons, Kate Jackson, Jonathan P. Weiner, and Brendan Saloner. A predictive risk model for nonfatal opioid overdose in a statewide population of buprenorphine patients. *Drug and Alcohol Dependence*, 201:127–133, 2019.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

Guang Cheng, Jingui Xie, and Zhichao Zheng. Optimal stopping for medical treatment with predictive information. *SSRN preprint 3397530*, 2019.

Roger Chou, Gilbert J. Fanciullo, Perry G. Fine, Jeremy A. Adler, Jane C. Ballantyne, Pamela Davies, Marilee I. Donovan, David A. Fishbain, Kathy M. Foley, Jeffrey Fudin, Aaron M. Gilson, Alexander Kelter, Alexander Mauskop, Patrick G. O'Connor, Steven D. Passik, Gavril W. Pasternak, Russell K. Portenoy, Ben A. Rich, Richard G. Roberts, Knox H. Todd, and Christine Miaskowski. Clinical Guidelines for the Use of Chronic Opioid Therapy in Chronic Noncancer Pain. *Journal of Pain*, 10(2):113–130, 2009.

Chronic Condition Data Warehouse. CCW Condition Algorithms (rev. 06/2019). Technical report, Centers for Medicare & Medicaid Services, 2019. URL https://www2.ccwdata.org/documents/10280/19139421/other-condition-algorithms.pdf.

Bryan N. Cochran, Annesa Flentje, Nicholas C. Heck, Jill Van Den Bos, Dan Perlman, Jorge Torres, Robert Valuck, and Jean Carter. Factors predicting development of opioid use disorders among individuals who receive an initial opioid prescription: Mathematical modeling using a database of commercially-insured individuals. *Drug and Alcohol Dependence*, 138:202–208, 2014.

Amber Cragg, Jeffrey P. Hau, Stephanie A. Woo, Sophie A. Kitchen, Christine Liu, Mary M. Doyle-Waters, and Corinne M. Hohl. Risk Factors for Misuse of Prescribed Opioids: A Systematic Review and Meta-Analysis. *Annals of Emergency Medicine*, 74(5):634–646, 2019.

Kraig Delana, Nicos Savva, and Tolga Tezcan. Proactive customer service: operational benefits and economic frictions. *Manufacturing & Service Operations Management*, 23(1):70–87, 2021.

Xinyu Dong, Sina Rashidian, Yu Wang, Janos Hajagos, Xia Zhao, Richard N. Rosenthal, Jun Kong, Mary Saltz, Joel Saltz, and Fusheng Wang. Machine learning based opioid overdose prediction using electronic health records. In *AMIA Annual Symposium Proceedings*, pages 389–398. American Medical Informatics Association, 2019.

Deborah Dowell, Tamara M. Haegerich, and Roger Chou. CDC Guideline for Prescribing Opioids for Chronic Pain—United States, 2016. *JAMA*, 315(15):1624–1645, 2016.

Robert Dufour, Jack Mardekian, Margaret K. Pasquale, David Schaaf, George A. Andrews, and Nick C. Patel. Understanding predictors of opioid abuse: predictive model development and validation. *American Journal of Pharmacy Benefits*, 6(5): 208–216, 2014.

Randall J Ellis, Zichen Wang, Nicholas Genes, and Avi Ma'ayan. Predicting opioid dependence from electronic health records with machine learning. *BioData Mining*, 12(1):3, 2019.

A. Fareed, S. Stout, J. Casarella, S. Vayalapalli, J. Cox, and K. Drexler. Illicit opioid intoxication: diagnosis and treatment. *Substance Abuse: Research and Treatment*, 5:SART–S7090, 2011.

Lindsey M. Ferris, Brendan Saloner, Noa Krawczyk, Kristin E. Schneider, Molly P. Jarman, Kate Jackson, B. Casey Lyons, Matthew D. Eisenberg, Tom M. Richards, Klaus W. Lemke, et al. Predicting opioid overdose deaths using prescription drug monitoring program data. *American Journal of Preventive Medicine*, 57(6):e211–e217, 2019.

Scott M. Fishman, Barth Wilsey, Jane Yang, Gary M. Reisfield, Tara B. Bandman, and David Borsook. Adherence monitoring and drug surveillance in chronic opioid therapy. *Journal of Pain and Symptom Management*, 20(4):293–307, 2000.

Curtis Florence, Feijun Luo, and Ketra Rice. The economic burden of opioid use disorder and fatal opioid overdose in the united states, 2017. *Drug and Alcohol Dependence*, 218:108350, 2021.

Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4 (Dec.):933–969, 2003.

Renu K. Garg, Deborah Fulton-Kehoe, and Gary M. Franklin. Patterns of opioid use and risk of opioid overdose death among Medicaid patients. *Medical Care*, 55(7): 661–668, 2017.

Peter Geissert, Sara Hallvik, Joshua Van Otterloo, Nicole O'Kane, Lindsey Alley, Jody Carson, Gillian Leichtling, Christi Hildebran III, Wayne Wakeland, and Richard A Deyo. High risk prescribing and opioid overdose: Prospects for prescription drug monitoring program based proactive alerts. *Pain*, 159(1):150, 2018.

Jason M. Glanz, Komal J. Narwaney, Shane R. Mueller, Edward M. Gardner, Susan L. Calcaterra, Stanley Xu, Kristin Breslin, and Ingrid A. Binswanger. Prediction model for two-year risk of opioid overdose among patients prescribed chronic opioid therapy. *Journal of General Internal Medicine*, 33(10):1646–53, 2018.

Julien Grand-Clément, Carri W Chan, Vineet Goyal, and Gabriel Escobar. Robustness of proactive ICU transfer policies. *arXiv preprint arXiv:2002.06247*, 2020.

Md Mahmudul Hasan, Md. Noor-E-Alam, Mehul Rakeshkumar Patel, Alicia Sasser Modestino, Leon D. Sanchez, and Gary Young. A novel big data analytics framework to predict the risk of opioid use disorder. *arXiv preprint arXiv:1904.03524*, 2019.

Justine S. Hastings, Mark Howison, and Sarah E. Inman. Predicting high-risk opioid prescriptions before they are given. *Proceedings of the National Academy of Sciences*, 117(4):1917–1923, 2020.

Kevin C. Heslin, Pamela L. Owens, Zeynal Karaca, Marguerite L. Barrett, Brian J. Moore, and Anne Elixhauser. Trends in opioid-related inpatient stays shifted after the US transitioned to ICD-10-CM diagnosis coding in 2015. *Medical Care*, 55(11): 918–923, 2017.

Wenqi Hu, Carri W. Chan, José R. Zubizarreta, and Gabriel J. Escobar. An examination of early transfers to the ICU based on a physiologic risk score. *Manufacturing & Service Operations Management*, 20(3):531–549, 2018.

Yue Hu, Carri W. Chan, and Jing Dong. Optimal scheduling of proactive care with patient deterioration. *Management Science*, forthcoming.

Timothy R. Hylan, Michael Von Korff, Kathleen Saunders, Elizabeth Masters, Roy E. Palmer, David Carrell, David Cronkite, Jack Mardekian, and David Gross. Automated prediction of risk for problem opioid use in a primary care setting. *Journal of Pain*, 16(4):380–387, 2015.

Interpretable AI, LLC. Interpretable AI Documentation, 2020. URL `https://www.interpretable.ai`.

Julian Kurz and Richard Pibernik. Flexible capacity management with future information. *SSRN preprint 2863088*, 2016.

Sauchi Stephen Lee. Regularization in skewed binary classification. *Computational Statistics*, 14(2):277–292, 1999.

Wei-Hsuan Lo-Ciganic, James L. Huang, Hao H. Zhang, Jeremy C. Weiss, Yonghui Wu, C. Kent Kwoh, Julie M. Donohue, Gerald Cochran, Adam J. Gordon, Daniel C. Malone, Courtney C. Kuza, and Walid F. Gellad. Evaluation of machine-learning algorithms for predicting opioid overdose risk among Medicare beneficiaries with opioid prescriptions. *JAMA Network Open*, 2(3):e190968–e190968, 2019.

Wei-Hsuan Lo-Ciganic, James L Huang, Hao H Zhang, Jeremy C Weiss, C Kent Kwoh, Julie M Donohue, Adam J Gordon, Gerald Cochran, Daniel C Malone, Courtney C Kuza, and Walid F Gellad. Using machine learning to predict risk of incident opioid use disorder among fee-for-service Medicare beneficiaries: A prognostic study. *PLOS ONE*, 15(7):e0235981, 2020.

Maimuna S. Majumder and Sherri Rose. Health care claims data may be useful for COVID-19 research despite significant limitations. *Health Affairs Blog*, 2020. URL `https://www.healthaffairs.org/do/10.1377/hblog20201001.977332/full/`.

Giovanna Menardi and Nicola Torelli. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):92–122, 2014.

National Academies of Sciences, Engineering, and Medicine. *Pain Management and the Opioid Epidemic: Balancing Societal and Individual Benefits and Risks of Prescription Opioid Use.* National Academies Press, Washington, D.C., 2017.

National Center for Injury Prevention and Control. CDC compilation of benzodiazepines, muscle relaxants, stimulants, zolpidem, and opioid analgesics with oral morphine milligram equivalent conversion factors, 2018. URL `https://www.cdc.gov/drugoverdose/resources/data.html`. Centers for Disease Control and Prevention, Atlanta, GA.

National Committee for Quality Assurance. HEDIS 2019 medication list directory (MLD) of NDC codes, 2018. URL `https://www.ncqa.org/hedis/measures/hedis-2019-ndc-license/hedis-2019-final-ndc-lists/`.

National Research Council and Institute of Medicine. *Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities.* National Academies Press, Washington, D.C., 2009.

New York City Department of Health and Mental Hygiene. Unintentional drug poisoning (overdose) deaths involving opioids in New York City in 2018. *Epi Data Brief*, 116, 2019. URL `https://www1.nyc.gov/assets/doh/downloads/pdf/epi/databrief116.pdf`.

Teryl K. Nuckols, Laura Anderson, Ioana Popescu, Allison L. Diamant, Brian Doyle, Paul Di Capua, and Roger Chou. Opioid prescribing: A systematic review and critical appraisal of guidelines for chronic pain. *Annals of Internal Medicine*, 160 (1):38–47, 2014.

Office of Inspector General. Toolkit: Using data analysis to calculate opioid levels and identify patients at risk of misuse or overdose. Technical Report OEI-02-17-00560, U.S. Department of Health & Human Services, June 2018. URL `https://oig.hhs.gov/oei/reports/oei-02-17-00560.pdf`.

Tae Woo Park, Lewei Allison Lin, Avinash Hosanagar, Amanda Kogowski, Katie Paige, and Amy S. B. Bohnert. Understanding risk factors for opioid overdose in clinical populations to inform treatment and policy. *Journal of Addiction Medicine*, 10(6):369–381, 2016.

Leonard J. Paulozzi, Edwin M. Kilbourne, Nina G. Shah, Kurt B. Nolte, Hema A. Desai, Michael G. Landen, William Harvey, and Larry D. Loring. A history of being prescribed controlled substances and risk of drug overdose death. *Pain Medicine*, 13(1):87–95, 2012.

Jordan S. Peck, James C. Benneyan, Deborah J. Nightingale, and Stephan A. Gaehde. Predicting emergency department inpatient admissions to improve same-day patient flow. *Academic Emergency Medicine*, 19(9):E1045–E1054, 2012.

J. Bradford Rice, Alan G. White, Howard G. Birnbaum, Matt Schiller, David A. Brown, and Carl L. Roland. A model to identify patients at risk for prescription opioid abuse, dependence, and misuse. *Pain Medicine*, 13(9):1162–1173, 2012.

Brendan Saloner, Hsien-Yen Chang, Noa Krawczyk, Lindsey Ferris, Matthew Eisenberg, Thomas Richards, Klaus Lemke, Kristin E. Schneider, Michael Baier, and Jonathan P. Weiner. Predictive modeling of opioid overdose using linked statewide medical and criminal justice data. *JAMA Psychiatry*, 77(11):1155–1162, 2020.

Substance Abuse and Mental Health Services Administration. Key substance use and mental health indicators in the United States: Results from the 2018 National Survey on Drug Use and Health. Technical Report PEP19-5068, NSDUH Series H-54, Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, Rockville, MD, August 2019. URL `https://www.samhsa.gov/data/`.

Yanmin Sun, Andrew K. C. Wong, and Mohamed S. Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4):687–719, 2009.

Ralph E. Tarter, Levent Kirisci, Gerald Cochran, Amy Seybert, Maureen Reynolds, and Michael Vanyukov. Forecasting opioid use disorder at 25 years of age in 16-year-old adolescents. *Journal of Pediatrics*, 225:207–213.e1, 2020.

United States Drug Enforcement Administration. Drug Scheduling, n.d. URL `https://www.dea.gov/drug-scheduling`.

Lynn R. Webster. Risk factors for opioid-use disorder and overdose. *Anesthesia & Analgesia*, 125(5):1741–1748, 2017.

Lynn R. Webster and Rebecca M. Webster. Predicting aberrant behaviors in opioid-treated patients: preliminary validation of the Opioid Risk Tool. *Pain Medicine*, 6 (6):432–442, 2005.

Lynn R. Webster, Susan Cochella, Nabarun Dasgupta, Keri L. Fakata, Perry G. Fine, Scott M. Fishman, Todd Grey, Erin M. Johnson, Lewis K. Lee, Steven D. Passik, John Peppin, Christina A. Porucznik, Albert Ray, Sidney H. Schnoll, Richard L. Stieg, and Wayne Wakeland. An analysis of the root causes for opioid-related overdose deaths in the United States. *Pain Medicine*, 12(Supplement 2):S26–S35, 2011.

Alan G. White, Howard G. Birnbaum, Matt Schiller, Jackson Tang, and Nathaniel P. Katz. Analytic models to identify patients at risk for prescription opioid abuse. *American Journal of Managed Care*, 15(12):897–906, 2009.

Nana Wilson, Mbabazi Kariisa, Puja Seth, Herschel Smith, and Nicole L. Davis. Drug and opioid-involved overdose deaths — United States, 2017–2018. *Morbidity and Mortality Weekly Report*, 69(11):290–297, 2020.

Kuang Xu and Carri W. Chan. Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing & Service Operations Management*, 18(3):314–331, 2016.

Galit B. Yom-Tov, Liron Yedidsion, and Yueming Xie. An invitation control policy for proactive service systems: Balancing efficiency, value, and service level. *Manufacturing & Service Operations Management*, Articles in Advance, 2020.

Barbara Zedler, Lin Xie, Li Wang, Andrew Joyce, Catherine Vick, Janet Brigham, Furaha Kariburyo, Onur Baser, and Lenn Murrelle. Development of a risk index for serious prescription opioid-induced respiratory depression or overdose in Veterans' Health Administration patients. *Pain Medicine*, 16(8):1566–1579, 2015.

Barbara K. Zedler, William B. Saunders, Andrew R. Joyce, Catherine C. Vick, and E. Lenn Murrelle. Validation of a screening risk index for serious prescription opioid-induced respiratory depression or overdose in a US commercial health plan claims database. *Pain Medicine*, 19(1):68–78, 2018.

Shaoquan Zhang, Longbo Huang, Minghua Chen, and Xin Liu. Proactive serving decreases user delay exponentially: The light-tailed service time case. *IEEE/ACM Transactions on Networking*, 25(2):708–723, 2016.