Algorithmic Fairness in Predicting Opioid Use Disorder using Machine Learning*

Angela E. Kilby

Northeastern University

November 2020

Abstract

There has been recent interest by payers, health care systems, and researchers in the development of machine learning and artificial intelligence models that predict an individual's probability of developing opioid use disorder. The scores generated by these algorithms can be used by physicians to tailor the prescribing of opioids for the treatment of pain, reducing or foregoing prescribing to individuals deemed to be at high risk, or increasing prescribing for patients deemed to be at low risk. This paper constructs a machine learning algorithm to predict opioid use disorder risk using commercially available claims data similar to those utilized in the development of proprietary opioid use disorder prediction algorithms. We study risk scores generated by the machine learning model in a setting with quasi-experimental variation in the likelihood that doctors prescribe opioids, generated by changes in the legal structure for monitoring physician prescribing. We find that machine-predicted risk scores do not appear to correlate at all with the individual-specific heterogeneous treatment effect of receiving opioids. The paper identifies a new source of algorithmic unfairness in machine learning applications for health care and precision medicine, arising from the researcher's choice of objective function. While precision medicine should guide physician treatment decisions based on the heterogeneous causal impact of a course of treatment for an individual, allocating treatments to individuals receiving the most benefit and recommending caution for those most likely to experience harmful side effects, ML models in health care are often trained on proxies like individual baseline risk, and are not necessarily informative in deciding who would most benefit, or be harmed, by a course of treatment.

*Department of Economics, Northeastern University. Email: a.kilby@northeastern.edu. Thanks to Alicia Sasser-Modestino, Bob Triest, Donghee Jo, Peter Hull, Byron Wallace, Leo Beletsky, seminar participants at Bentley University, and others, including the participants of the Northeastern Faculty Works-in-Progress Colloquium, for feedback. Funding support for an early phase of this project is gratefully acknowledged from the National Institute on Aging Post-Doctoral Fellowship and the National Science Foundation Graduate Research Fellowship.

1 Introduction

Since 1999, the opioid crisis has claimed between half a million and three-quarters of a million lives in the United States (Hedegaard, Miniño, and Warner, 2020).[1] As the crisis has accelerated, a key focus of researchers and legislators has been to determine the appropriate role of opioid pharmacotherapy in the treatment of pain conditions.

On the one hand, the clinical value of opioids prescribed for acute, chronic, and cancer pain has been well-established by many studies (Chaparro et al., 2014; Furlan et al., 2006) – though some studies have questioned whether that value has been is overstated for chronic, non-cancer pain (Chou et al., 2009; Dowell, Haegerich, and Chou, 2016). Denying opioid therapy to pain patients with no other viable options for pain management can have profound negative consequences for their physical and mental health that arises from their untreated pain (Dueñas et al., 2016).

On the other hand, many risks accrue to individuals and society from the practice of prescribing opioids liberally for pain management. Patients may, in the course of opioid pharmacotherapy, develop opioid use disorder (OUD). Additionally, opioids may be diverted from the medical system – taken out of the medicine cabinets of relatives, prescribed to all comers by unscrupulous physicians operating "pill mills," or stolen for sale on the street in pharmacy thefts. At the societal scale, levels of opioid prescribing are correlated to the magnitude of opioid-related harms. As prescribing rose threefold between 2001 and 2011, opioid overdose deaths also rose threefold, as can be seen in Figure 1. However, they are not the only driver: opioid overdose deaths continued to climb after opioid prescribing began to fall in 2012. Deaths linked specifically to heroin, an illicit opiate, overtook those linked to prescription opioids in mid-2015, and those involving fentanyl, a synthetic opioid commonly added to black-market opioids, overtook heroin shortly thereafter in early 2016.[2]

For these reasons, tensions have arisen as physicians have attempted to codify best practices into medical guidelines and as legislators have endeavored to pass opioid prescribing legislation (Alford, 2016; Nicholson, Hoffman, and Kollas, 2018).

This core tradeoff – a desire to address overprescribing that has led to rising rates of opioid use

[1] From 1999-2018, the CDC recorded 769,935 drug overdose deaths, of which 446,032 were specifically linked to prescription opioids or heroin; because not every overdose death certificate specifically identifies the contributing drugs, the opioid-specific count is a lower bound (Ruhm, 2018).

[2] A growing body of research attributes these changing dynamics of mortality and rising overall lethality of the opioid crisis in recent years to increasingly restrictive prescribing practices beginning in the mid-2000s. These restrictions led to scarcity of illicitly-obtained prescription opioids, which can be dangerous or lethal to abuse, but are still safer than heroin due to their manufactured nature, which yields a predictable dose with no adulterants. As prescribing restrictions led to a shift away from black market opioid pills and towards heroin (increasingly frequently adulterated with fentanyl) and counterfeit opioid pills containing synthetic fentanyl, the crisis grew more lethal (Evans, Lieber, and Power, 2019).

disorder and overdoses, while not inadvertently harming pain patients in need of opioid therapy – has led to a search for solutions that enable doctors, pharmacists, and insurers to prescribe opioids in a smarter and more tailored fashion (Bruehl et al., 2013). There has been particular recent enthusiasm for new Artificial Intelligence and Machine Learning (AI/ML) models, which use varied, new, and extremely rich sources of patient-level health data to predict an individual's opioid use disorder risk. These models are designed to support prescriber decision-making, helping to tailor opioid pain pharmacotherapy based on machine assessment of the patient's risk profile, and have already been deployed inside several state Prescription Drug Monitoring Programs (Appriss Health, 2018; Speights and Atencio, 2018). Physicians, providers, and pharmacists are generally expected to consult the patient's record inside the PDMP database prior to prescribing or dispensing scheduled drugs such as opioids, and in states adopting one proprietary algorithm the risk score is displayed prominently on the patient's record (Ohio Board of Pharmacy, 2019).

While such algorithms deployed in a clinical setting are appealing for their potential ability to resolve tradeoffs inherent in opioid prescribing, the stakes of making treatment decisions based on such an algorithm are high: a patient incorrectly categorized as "low-risk" and prescribed an opioid could subsequently develop debilitating opioid use disorder, and a patient incorrectly categorized as "high-risk" could have needed pain treatment withheld.

In this paper we identify a new source of "algorithmic unfairness" in the typical development of risk-prediction algorithms intended for use in a health care setting. While researchers and companies working commercial algorithms typically build ML models to identify patients at *high baseline risk*, assuming that averting prescribing (or otherwise changing the course of treatment) to patients at high overall risk will have a differentially large causal impact on the development of future opioid use disorder, we argue that this objective function is incorrectly chosen for the task of identifying *patients who would most benefit from having opioid prescribing averted.* The typical objective function for the machine learning task is to identify patients most likely to develop OUD; the correct objective function identifies patients with the most negative heterogeneous causal treatment effect, i.e., patients whose chances of developing OUD become markedly greater after having been prescribed an opioid, and conversely, patients for whom receipt of an opioid has little effect on subsequent development of OUD. Estimating these heterogeneous treatment effects is likely to be especially challenging when the assignment to treatment is confounded with other patient characteristics that might influence the probability of developing opioid use disorder (Kleinberg et al., 2015; Lada et al., 2019) .

To investigate the clinical value and fairness of these ML models, we first successfully train a machine learning model to predict opioid use disorder in keeping with other opioid risk prediction models in academic literature and in commercial development. We consider the theoretical conditions in which typical observational ML approaches will uncover the heterogeneous treatment effects of interest. We then test those conditions by nesting the risk predictions generated by this model in a quasi-experimental setting where opioid prescribing was reduced across-the-board, finding that the heterogeneous treatment effect of reducing opioid prescribing on opioid use disorder is uncorrelated with the risk score generated by the machine. In other words, models trained with the typical risk-prediction objective function do not produce a valid proxy for the object of interest, patient-level heterogeneous treatment effects.

We find that the machine identifies high risk for opioid use disorder based on a few key demographic characteristics, as well as flagging complex chronic pain patients with a number of comorbidities as high risk, but these patients do not on average benefit from a reduction in prescribing more than any other group. In fact, results suggest that reallocating prescribing according to machine recommendation, in a quantity-neutral manner, away from groups with high risk scores and towards groups with low risk scores, might paradoxically *increase* the prevalence of opioid use disorder.

1.1 Related Work

This work bridges the gap between the machine learning literature, where ML algorithms are in active development for medical applications, and the burgeoning literature on heterogeneous causal treatment effects.

There have been a number of recent academic and commercial efforts to develop practically-oriented ML algorithms that can risk-stratify patients to inform care decisions (Wiens et al., 2019). An overview of work in this area specifically related to opioid risks and prescribing is found below in Section 2. Similar efforts have applied machine learning techniques to predict overall risk of deteriorating health in order to target at-need patients with increased services (Obermeyer et al., 2019), to predict which discharged patients are likely to be readmitted (Morgan et al., 2019), and to identify patients at greatest risk of hospital-acquired infection (Oh et al., 2018).

There is also a burgeoning econometric and computational literature that seeks to incorporate machine learning techniques to directly estimate treatment effect heterogeneity (Athey, Tibshirani, and Wager, 2019; Athey and Wager, 2019; Lada et al., 2019; Powers et al., 2018; Wager and Athey, 2018). These papers seek to develop causal methods that can answer counterfactual questions about what

would happen to individuals if they receive treatment A or treatment B, and specifically stratify these treatment effects according to observable individual covariates that reliably predict the sign and magnitude of the individual's treatment effect. One of many compelling applications for these techniques is in the development of "personalized medicine" – specifically, algorithms that allow the personalization of medical care decisions with treatment choices that have the greatest positive impact given the individual's characteristics. This paper provides a theoretical and empirical bridge between these two literatures, and demonstrates the need for the advancement of techniques to estimate heterogeneous treatment effects directly.

The paper proceeds as follows. In Section 2.1 we detail the construction of a machine learning model to predict an individual's probability of developing opioid use disorder, following methods standard in this literature. In Section 2.2 we discuss its performance according to traditional metrics of ML model performance. In Section 3 we consider the theoretical link between a risk prediction model and heterogeneous treatment effects. In Section 4 we nest this model in a quasi-experimental setting where opioid prescribing was reduced, to investigate whether risk strata generated by the model correlate to heterogeneity in outcomes. Section 5 concludes with a discussion of the potential ramifications of incorporating these models into clinical decision-making.

2 Algorithms to Identify Opioid Risk

The growing harms of the opioid crisis have made clinical decisions around the prescribing of opioids for pain increasingly frought. Dual goals of providing opioid pain relief to patients in need, while not inadvertently causing harm, have spurred the development of a variety of approaches to mitigate this risk, including increasingly detailed prescribing guidelines and increasingly restrictive legislation (Dowell, Haegerich, and Chou, 2016). An area in active development has been algorithms and models that identify specific patients at elevated risk of developing opioid use disorder, so that opioid prescriptions can be averted for those at high risk and allowed for those at low risk. First-generation efforts to identify at-risk patients used decision-support algorithms based on a small number of "red flag" characteristics, including number of prescriptions, number of prescribers, and number of pharmacies visited (Cepeda et al., 2012; Katz et al., 2010; Parente et al., 2004; Sullivan et al., 2010). Second-generation efforts have made significant technical advances, applying machine learning techniques to data with much larger numbers of clinical covariates. These efforts generally use clinical covariates that can be found in commercial medical insurance claims, electronic medical records, and other sources of "big health

data," and generate a continuous risk score for developing opioid use disorder that can be exposed to clinicians or payers for use in decision-making. (Brenton et al., 2017; Che et al., 2017; Hasan et al., 2019; Hastings, Howison, and Inman, 2020; Lo-Ciganic et al., 2019).

These models have been developed by academic researchers, as well as commercially. A primary target for deployment of commercial versions of these algorithms is inside Prescription Drug Monitoring programs, which are databases that collect records on patients' controlled substance prescribing history, and that doctors and providers will consult before writing a prescription for an opioid. A leading PDMP provider recently began to offer a proprietary "overdose risk score," generated using machine learning, that can be displayed prominently at the top of the patient's record (Appriss Health, 2018; Speights and Atencio, 2018); this score is currently deployed in a handful of states. An earlier version of the score, the NarxCare proprietary score, which was developed using the "red flag" approach, is integrated into PDMP in 20 states and into most major pharmacy chains nationwide, including Walmart, CVS, Walgreens, Kroger, and Rite Aid. LexisNexis, Milliman, HBI Solutions, and others also market proprietary machine learning/AI tools to predict opioid risk to health insurers and providers (Ravindranath, 2019). Algorithms are typically trained on health insurance claims data, and sometimes bring in nontraditional sources of data (social networks, housing status, income).

2.1 Model Construction

We train a machine learning model to predict risk of opioid use disorder that shares key characteristics with the above literature. We use a common source of medical data, commercial health insurance claims data, and follow the above literature for guidance on label, feature, and cohort selection. Per that literature, we build an ML model, trained on a cohort of patients receiving a new opioid prescription who were previously opioid-naive, that identifies factors correlated to eventual development of OUD after that initial opioid prescription.

2.1.1 Data

The data used to train the machine learning model are extracted from commercial health insurance claims data: Truven Marketscan covering years 2005-2012. These data are sourced from large employers, and for the years of this study, cover approximately 7 million individuals representing 175 million employed Americans with employer sponsored health insurance. The data include all insurance claims for services provided, including prescription drug claims filled, and inpatient and outpatient utilization,

and contain rich diagnosis and procedure data as a part of each claim. Individual enrollee identifiers and date of service indicators allow for the construction of panel data at the individual-month level for clinical variables of interest.

One weakness of this data source is that it is only representative of individuals with private employer-sponsored insurance, and thus models trained on this data will not necessarily be externally generalizable to critical populations who might be most at risk of opioid use disorder, such as the uninsured, unemployed, or those on Medicaid. This weakness is shared with most academic and commercial efforts in this space, which often rely on claims data because it is readily available. While some of the resulting algorithms are intended for use on a similar insured population, others are intended to be deployed more broadly.

2.1.2 Labels

The outcome variable of interest is the development of opioid use disorder, encoded as a binary classifier. Following other work, we identify potential opioid use disorder using opioid dependence and opioid overdose diagnosis codes, as well as observed prescriptions for OUD recovery medications, and claims in mental health and substance abuse facilities. Specifically, we use a custom episode grouper to link related claims across time, and consider a person to have developed OUD if they meet two out of the following three criteria in the same month:[3]

1. Has any ICD-9 diagnosis code for opioid dependence: 304.00-304.03, 304.70-304.73, 305.50-305.53, 965.00-965.02, 965.09, E850.0-E850.2;

2. Has a recovery prescription, identified by any prescription claim containing buprenorphine as an active ingredient (e.g., Suboxone/Subutex, a medication to treat opioid use disorder);

3. Claim identified as occuring in a mental health or substance abuse facility, or with a mental health or substance abuse provider (see Appendix Figure A.1).

This procedure identifies 22,655 enrollees (out of the ∼7 million total enrollees) as having OUD. If the sample were nationally representative, this rate of 3 cases per 1000 would imply ∼941,000 Americans with OUD, which is lower than the estimated 1.8 million Americans with OUD in that time period. This undercount reflects widespread under-diagnosis and under-treatment of opioid use disorder in general, and is in line with rates of case identification in other work (Barocas et al., 2018).

[3]Code for the custom episode grouper is available on request.

The under-identification of cases means many positive cases occuring among individuals in our sample may be labeled falsely as negatives, potentially introducing bias in the estimation of the ML algorithm. The validity of the resulting model requires the assumption that those identified as having OUD in insurance and health care data are broadly representative of the uncounted OUD sufferers, and specifically that the presence or absence of an observable diagnosis for those with OUD will be uncorrelated with the covariates used to train the model.

However, there are many reasons to be concerned that these assumptions may not hold. Receiving a formal opioid use disorder diagnosis will be in part a function of an individual's interaction with the health care system, which will in turn be related to socioeconomic factors. Further, individuals in frequent contact with medical providers, for example those receiving opioid pharmacotherapy for pain conditions, are more likely to be evaluated for opioid use disorder, especially in an environment of increasing scrutiny on pain patients, where clinical differentiation between normal dependence and problematic use is somewhat subjective and standards have shifted. This deficiency in labels that are extracted from medical claims and electronic medical records will be a concern in all ML models designed to predict opioid use disorder.

2.1.3 Features

Covariates (model features) are extracted from claims data records, and indicators of clinical experience are transformed to a panel data structure with observations at the individual-month level. Covariates used include sex, 5-year age bins (Age 26-30, etc.), dummy variables for industry of employment, and a large panel of clinical variables, including dummy indicators for 3-digit ICD-9 codes, facility types, and provider types. In total, 1,371 features were included in the baseline model.

2.1.4 Cohort Selection

We select a cohort of opioid-naive patients with no prior opioid use disorder diagnosis; this approach follows Hastings, Howison, and Inman (2020) and others. The selection procedure is as follows:

1. Select all patients with any opioid prescription claim; the date of first prescription is their index date (7,385,007 -> 919,573)

2. Drop any patients without 1 year of claims eligibility prior to index date (919,573 -> 553,183)

3. Drop any patients with an OUD episode that begins prior to the index date (553,183 -> 550,636)

4. In the panel data, an observation at the individual-month is assigned a positive label if it is followed by an OUD episode within five years

In step 3 of the above selection procedure, for enrollees that have both OUD and an opioid prescription during their period in the claims data, 44% are prescribed an opioid prior to their first OUD episode; but 56% have an OUD diagnosis that precedes their initial opioid prescription. Thus over half of OUD sufferers in the data are dropped as ineligible.[4]

2.1.5 Algorithm

This is a binary classification machine learning exercise, and we use an off-the-shelf machine learning algorithm, gradient boosted decision trees, implemented in Python by the XGBoost library. The algorithm, which is an ensemble method, fits a large number of trees sequentially (each tree allowing for a non-linear relationship structure between covariates) with iterative improvements each time (Friedman, Hastie, and Tibshirani, 2009). The resulting model generates a score at the individual-month level that reflects the machine-determined probability of observing an OUD diagnosis in that month; in a given month, if an individual has features that are correlated with OUD, they will receive a higher score than individuals who do not. We determine an individual's overall risk of OUD according to their maximum machine-predicted score over their period in the sample. We evaluate model performance at the individual level, on a 10% hold-out sample of enrollees. An individual is deemed to be a true positive at a given threshold if their risk score is above that threshold and they have ever received an OUD diagnosis, and a false positive if they have never reeceived one.

2.2 Model Results

2.2.1 Feature Importance

An advantage of using a tree-based machine ensemble model such as gradient boosted decision trees is the ability to inspect the most important features that the model uses for prediction, allowing some visibility into the black box of the algorithm and a degree of interpretability. Figure 2 displays the most important features used by the trained model. "Total gain" captures the improvement in accuracy a feature contributed when added to a decision tree, averaged over all trees in the ensemble, and the

[4]This reveals a surprising absence of a temporal relationship between opioid prescriptions and opioid use disorder diagnoses; a given OUD sufferer is as likely to have had their OUD diagnisis precede their first opioid prescription as follow it. The absence of the expected temporal relationship calls into question the causality of the relationship, and provides informal evidence that a model built to predict opioid use disorder may not uncover a causal relationship with initial opioid prescription that can be used to guide prescribing and avert future diagnoses.

reported total gain thus represents the relative importance of each feature in predicting an individual's risk of developing opioid use disorder in the final ensemble model.[5]

Several features of interest stand out. First, a number of demographic indicators appear, including indicators for a young age (Age 26-31, Age 31-36), male, and having blue collar employment (Industry: Manufacturing and Industry: Transportation, Communications, and Utilities).

Second, various diagnosis and procedure codes that indicate complex pain conditions play a prominent role. Near the top of the list is the provider code for chiropractor, and high on the list are diagnoses for intervertebral disc disorders, other and unspecified disorders of the back, nonallopathic lesions, sprains and strains of the ankle and foot, and other disorders of the cervical region.

There are also indicators of mental health comorbidities: sleep disorders, anxiety and dissociative disorders and an inpatient psychiatric facility.

Finally, there are several potential indicators of undiagnosed opioid use disorder: viral hepatatis, diseases of the nail, and "drug dependence" (drug dependence that is not dependence on opioids).

2.2.2 Model Performance

A typical measure of machine learning model performance on a binary classification problem is the receiver operating characteristic (ROC) curve and area under the curve (AUC) statistic. As shown in the left-hand panel of Figure 3, on the AUC metric, the model performs well, with an AUC of 0.890 in the hold-out test sample. Additionally, we check its usefulness on predicting opioid use disorder in the sample of individuals that were excluded from the study because they did not meet the cohort selection criteria; the AUC for those individuals is 0.955.

However, because the sample is highly unbalanced– the ratio of positive labels to negative labels is 0.0038 – this performance metric obscures significant concerns with precision, as can be seen in the right-hand panel of Figure 3. In particular, any given threshold chosen for implementing the model will include many false positives, or individuals labeled as at-risk for opioid use disorder (who would be targeted for prescription reduction) that do not ever develop OUD, or at least not that was observed in the data.

Specifically, choosing a cutoff threshold set at the 99th percentile of the score distribution would yield 7,640 true positives and 61,255 false positives (precision = 11%). 13,231 positives would additionally be misclassified as false negatives (for a recall, or a true positive rate, of 37%) A threshold at the

[5]Note that the sign of the relationship between the feature and opioid use disorder is not reported, only whether it is important to improving model accuracy.

95th percentile would correctly identify most positives, but have a very large number of false positives: it identifies 15,619 true positives and 328,715 false positives (precision = 4.5%), and 5,252 positives would be misclassified as false negatives (recall = 75%). For both thresholds, and any other reasonable threshold, the vast majority of enrollees (6.5 million or more) are classified correctly as true negatives (false positive rate of <1% for 99th percentile threshold, e.g. specificity of 99%).

The 95th and 99th percentile thresholds are common in literature for predicting opioid use disorder: guidance to prescribers for the commercial NarxScore and Overdose Risk Score specifically calls out the clinical utility of thresholds set those two levels. Other literature highlights other metrics such as the odds ratio; the algorithm does a good job of stratifying risk by these traditional metrics as well: the odds ratio of developing OUD in the 10th decile is 500x that in the 1st decile.

A recent review of five OUD prediction algorithms by Rough et al. (2019) found similar statistical characteristics as the model developed here: the precision of various models ranged from 1-15%; sensitivity from 1.4% to 15.3%; specificity from 90-100%.

3 Conceptual framework: Risk scores, heterogeneous treatment effects, and clinical utility

We first consider how observational data including treatment decisions, patient characteristics, and outcomes can be used to estimate heterogeneous effects. Consider a simple model of the effects of treatment by opioid therapy on the outcome of opioid use disorder:

$$Y_i(D_i) = \mu_i D_i + a_i \tag{1}$$

Here, $Y_i(1)$ is the potential opioid use disorder outcomes after receiving an opioid, $Y_i(0)$ is the potential opioid use disorder outcome having not received an opioid. $a_i = f(\mathbf{X}_i)$ is the patient-specific baseline opioid use disorder outcomes as a function of characteristics, and $\mu_i = g(\mathbf{X}_i)$ is the patient-specific increase in likelihood of developing OUD after being prescribed an opioid (according to D_i), i.e., the heterogeneity in treatment effect, which is the object we want to estimate.

We may make an unconfoundedness assumption,

$$D_i \perp\!\!\!\perp Y_i(0), Y_i(1) \mid \mathbf{X_i} \tag{2}$$

which would be satisfied in a randomized experiment or in some observational settings where treatment is not dependent on potential outcomes after conditioning on all observed covariates. If unconfoundedness holds, μ_i can be estimated as follows:

$$\mu_i = \mathbb{E}[Y_i \,|\, D_i = 1,\, \mathbf{X_i} = x] - \mathbb{E}[Y_i \,|\, D_i = 0,\, \mathbf{X_i} = x] \tag{3}$$

However, if unconfoundedness does not hold, then there may be selection/omitted variables bias in estimation of μ_i using standard techniques, where:

$$\begin{aligned} \mathbb{E}[Y_i \,|\, D_i = 1,\, \mathbf{X_i} = x] - \mathbb{E}[Y_i \,|\, D_i = 0,\, \mathbf{X_i} = x] = \\ \mathbb{E}[Y_{1i} \,|\, D_i = 1,\, \mathbf{X_i} = x] - \mathbb{E}[Y_{0i} \,|\, D_i = 1,\, \mathbf{X_i} = x] \\ + \{\mathbb{E}[Y_{0i} \,|\, D_i = 1,\, \mathbf{X_i} = x] - \mathbb{E}[Y_{0i} \,|\, D_i = 0,\, \mathbf{X_i} = x]\} \end{aligned} \tag{4}$$

The first term is the average treatment effect on the treated, and the second term represents selection bias in who receives treatment and who does not.

We may think selection may be a severe problem in this domain. Even conditioning on observable clinical covariates that can be found in the rich clinical data used in this literature, assignment to opioid therapy is likely to be endogenous to unobservable characteristics that may influence outcomes. In other work, we show that assigning opioid therapy quasi-experimentally reverses the sign of the impact of opioid therapy on key patient outcomes, relative to estimating those associations only using observational data (Kilby, 2019).

As described above in Section 2.1.4, the procedure used to construct our ML model, which mirrors that used in academic and commercial approaches to building opioid risk tools, addresses this selection problem by focusing on estimating patient risk conditional on having received treatment, shutting down the channel by which selection might contaminate estimates of μ_i. Instead, the models proxy for μ_i by estimating patient-specific increase in likelihood of developing OUD after being prescribed an opioid:

$$\mathbb{E}[Y_{1i} \,|\, D_i = 1,\, \mathbf{X_i} = x] = \mathbb{E}[Y_{1i} \,|\, D_i = 1,\, \mathbf{X_i} = x] = \mathbb{E}[\mu_{1i} + a_{1i} \,|\, \mathbf{X_i} = x] \tag{5}$$

This procedure to estimate $\tilde{\mu}_i$ will provide a rank-ordering of patients according to their heterogeneous treatment effects, μ_i, as long as baseline opioid use disorder risk $a_i = f(\mathbf{X}_i)$ is sufficiently correlated with the hetergeneous treatment effect $\mu_i = g(\mathbf{X}_i)$, or in other words, if $\mu_i + a_i$ is a good proxy for μ_i, which occurs when $Corr(\mu_i, a_i) >> 0$. However, it is not clear that there is an a priori reason

to assume baseline risk and treatment effect heterogeneity will have the same relationship to patient characteristics, or even have the same sign. Below in our quasi-experiment, we will test this assumption empirically.

3.1 Clinical value and welfare

We also consider a simple welfare model of the costs and benefits of implementation of these algorithms into clinical practice, following Hastings, Howison, and Inman (2020). They focus on the predictive accuracy of their model, and specifically, its True Positive (TP) and False Positive (FP) rates, in identifying who should receive or be denied an opioid, and present a condition for when the costs of implementing the model into clinical practice are outweighed the benefits.

$$TP(C_A - C_H) - FP * C_H > 0 \implies \frac{TP}{FP + TP} > \frac{C_H}{C_A} \tag{6}$$

Here, C_A is costs to the individual and society of an opioid-related adverse outcome, and C_H are costs to an individual of not receiving an opioid to treat pain. $\frac{TP}{FP+TP}$, or the percent of positives that are true positives, is precision.

Using True Positive and False Positive rates from Section 2.2.2 above, if True Positives are approximately 10% of all positives (as was the case at a 99% threshold; precision was approximately 11%), the costs of an individual having OUD must be 10X costs of not receiving an opioid to treat pain, in order for the implementation of the algorithm into clinical practice to be welfare-improving. Many papers in this literature, such as Hasan et al. (2019) are concerned with precision, and argue that C_H is very small and C_A is very large, i.e., there are no major costs associated with denying people opioids. Thus, even though it is common for precision to be low and false positive rates to be high, these models are found to be welfare-improving.

However, a less-discussed consideration is that this welfare model assumes every True Positive individual who is labeled high-risk and would have had an adverse outcome is prevented from having that adverse outcome by the implementation of the model into practice. In other words, the model assumes that the individual causal treatment effect is well-proxied for by the baseline risk levels as estimated in the standard ML approach, the assumption we will test in our quasi-experiment below.

4 Quasi-Experiment: State-level restrictions on opioid prescribing

4.1 Empirical Strategy

As highlighted in Sections 1 and 3 and in related work (Obermeyer et al., 2019), there is a conceptual gap between the goal of predicting baseline risk of OUD, and the identification of heterogeneous causal treatment effects. Putting risk-prediction algorithms into clinical practice to guide physician decision-making assumes that heterogeneous effects are monotonically increasing in the estimated baseline risk, but this assumption is not usually tested.

Here, we will use the predicted risk scores from the machine learning model trained above to stratify a patient population in a quasi-experimental setting where opioid prescribing was broadly reduced. If the machine-predicted risk scores are correlated with individual heterogeneous effects, and they are able to add clinical value above and beyond information physicians already have when deciding which patients to target for reduced prescribing, then the magnitude of treatment effect should be increasing in the risk score. This exercise is similar in spirit to Lada et al. (2019), where high-dimensional observational data was first used to estimate potential observational heterogeneous treatment effects in the value a user may receive from being recommended additional pages on Facebook, and those strata were then investigated inside an experimental setting where the treatment was randomized, to show the estimated heterogeneous effects positively correlated with heterogeneous causal treatment effects estimated experimentally.

The quasi-experimental setting we exploit is the timing of the initial establishment of statewide Prescription Drug Monitoring Programs (PDMPs). From 2004-2014, 38 states established PDMPs, which are databases that electronically track prescribing and dispensing of controlled prescriptions to patients (see Appendix Figure A.2). They are accessible to prescribers, dispensers, and law enforcement, and are intended for many purposes, including facilitating physician access to prescribing records in order to coordinate care across providers, prevent duplicative prescriptions, and allow the physician to scrutinize the patient record for evidence of bad behavior like doctor hopping/shopping. They are also used by authorities in most states to police or monitor the behavior of both patients *and* doctors, including surveillance to identify evidence of pill mills (bad prescriber behavior) and doctor shoppers (bad patient behavior) (U.S. Department of Justice, 2015).

In other work, we show that the timing of the introduction of PDMPs appears to have a broad,

causal, negative impact on opioid prescribing (Kilby, 2019), and in surveys doctors report that the introduction of PDMPs motivates a broad reduction in overall prescribing (Lin et al., 2017). This broad reduction motivates an analysis in the spirit of instrumented difference-in-differences (Hudson, Hull, and Liebersohn, 2017), where PDMPs generates quasi-experimental variation in levels of opioid prescribing. This quasi-experimental variation induced by a policy change allows investigation of the causal relationship between reduced opioid prescribing and health and other outcomes of interest.

We will study the impact of these PDMP laws on patients in the same commercial claims database (Truven Marketscan Commercial Claims and Encounters), stratified according to their predicted opioid risk score.

The econometric approach uses two related linear probability model specifications as follows:

$$Y_{it} = \alpha + \gamma_e + \lambda_t + \sum_{\tau} \sigma_\tau D^r_{\tau,st} + \beta X_{it} + \epsilon_{it} \tag{7}$$

$$Y_{it} = \alpha + \gamma_e + \lambda_t + \beta^r_1 \ 1(PDMP_{st}) + \beta_2 X_{it} + \epsilon_{it} \tag{8}$$

where i is individual in risk score strata r that entered the claims data sample with entry cohort e, s is state, and t is quarter. $D^r_{\tau,ist}$ are dummy variables for each quarter before and after the policy is introduced, estimated separately for each risk stratum, with τ normalized to 0 in the quarter physicians gain access to PDMP. Y_{it} is a binary outcome variable constructed from claims data, 1(received opioid) or 1(opioid abuse episode) in quarter. γ_e are entry-cohort fixed effects, and λ_t are time fixed effects.[6] X_{it} are individual controls: age, employment, and industry. Finally, robust standard errors, clustered at the state level, are reported in each specification.

Equation 7 allows for the estimation of leads and lags of the impact of the PDMP law changes, which can then be visualized in a graph. The leads can be used for a falsification test on the parallel trends identifying assumption by examining them for evidence of pre-trends in the adopting versus non-yet adopting states prior to PDMP implementation. The lags visualize any dynamic treatment effects that accrue in the years after the laws were passed. Equation 8 summarizes the treatment effect of the passage of PDMPs into one coefficient, $\hat{\beta}^r_1$.

[6]Our data are at the individual level, and individuals enter and exit the Truven sample in cohorts, usually based on whether their employer is selling their insurance claims data to the database that year. Because this individual-level panel is not stable in its composition, we use entry-cohort fixed effects here to control for the invariant characteristics of individuals; in Kilby (2019) we show that these entry-cohort fixed effects perform nearly identically to individual fixed effects, but have much reduced dimensionality, which is required for estimation of some specifications.

4.2 Results for full sample of all enrollees

We first show a "first stage:" that PDMPs affect overall prescribing on average for all enrollees, validating this setting as an appropriate quasi-experiment to investigate cohort-specific heterogeneous effects. In Figure 4 and Table 1, respectively, the impact of PDMP on overall opioid prescribing, estimated according to Equations 7 and 8 for one risk stratum (the full sample), is shown to be significantly negative and lasting. Leads are estimated to be small in magnitude and are statistically indistinguishable from zero, increasing confidence in our quasi-experimental setting.

Second, in Figure 5 and Table 2, we show that the implemention of a PDMP, and the reduction in opioid prescribing just shown in Figure 4, is associated with a concomitant reduction in observed opioid use disorder diagnoses across the full sample of enrollees, and that this reduction in OUD follows a similar overall pattern related to the timing of the implementatino of a PDMP as the reduction in prescribing.

Of note is our approach to quantify the magnitude of the estimated effect sizes. We will need an understanding of the effect size for each risk stratum, in order to assess whether heterogeneous effects are relatively large or small in each. We do this by calculating the "relative risk reduction," or relative decrease in the risk of an outcome, such as receiving an opioid, in the group treated by a PDMP, compared to the control group.[7] In the full sample, the probability of a given individual receiving an opioid in a given quarter is small: only 2.58% will be prescribed an opioid on average in the pre-period. As such, an estimated 0.0874 percentage point absolute drop after the implementation of a PDMP represents a relative risk reduction of being prescribed an opioid of approximately 3.49%, as can be seen in Table 1. The pre-period mean of developing OUD in a given quarter is even smaller in the full sample: 0.15%. But the estimated reduction of -0.014 yields a relatively substantial relative risk reduction of approximately 8.71%, as can be seen in Table 2.[8]

[7] Specifically, we calculate the relative risk reduction with standard errors from the regression postestimation of the linear probability model using `margins` for marginal effects and `nlcom` (non-linear combinations) in Stata to calculate $RRR \approx -\frac{\text{Estimated effect}}{\text{Pre-Mean}}$.

[8] The importance of working with relative risk reductions when examining heterogeneous treatment effects can be illustrated with a simple numerical example. Consider a group of 1,000 people, 100 of whom are prescribed an opioid. Suppose that every person receiving an opioid further has a 10% chance of developing opioid use disorder, meaning there will be 10 OUD cases, and that this risk is influenced by no individual characteristics other than the receipt of the opioid itself – this is thus a constant treatment effect scenario, with no heterogeneity in the impact of receiving an opioid on outcomes. Now, suppose an algorithm existed that separated the 1,000 people into two groups: one of 100 people, 5% of whom develop OUD, and one of 900 people, 0.555% of whom develop OUD. That implies 50 opioid-receivers are in group 1, and 50 opioid-receivers are in group 2, so 5 people who develop OUD are in each group as well. Consider now an experiment that cuts in half opioid prescribing, hitting all groups equally. In our linear probability model, the pre-period mean for group 1 of developing OUD is 0.05, and the estimated treatment effect of the intervention is $\hat{\beta}_1^1 = -0.025$. In group 2, the pre-period mean is 0.00555, and the estimated $\hat{\beta}_1^2 = -0.002775$. As is discussed in Harrell (2018), it is not uncommon for researchers to interpret the dramatically different $\hat{\beta}$s across the two groups, which are their absolute risk reductions, as evidence of heterogeneous treatment effects, but this is incorrect: the groups have constant effects. The

4.3 Results for two risk cohorts

Next, we stratify the sample into two groups: those above the 98.5th percentile of the machine-estimated risk score, and those below. We choose this threshold both because it is in keeping with those often used in clinical practice, as discussed above in Section 2.2.2, and also for analytical convenience: above and below this cutoff are approximately the same number of *OUD cases* observed. (See Footnote 8 above for a numerical example that illustrates this in detail.) The targeted (>98.5%) group is much smaller - 72,000 versus 4.9 million enrolllees - and thus has a much higher *rate* of OUD, but each group contains the same total number of cases.

As can be seen in Table 3, Columns (1) and (3), doctors do in fact reduce prescribing for the targeted group more than the non-targeted group: the appriximate relative risk reduction in opioid prescribing is 0.0587 in the targeted group versus 0.0304 in the non-targeted group. The algorithm and doctors both seem to believe that this group, characterized per Section 2.2.1 above as younger, male, and with chronic pain conditions, is at higher risk and in need of greater relative reductions after the introduction of a PDMP.

However, as can be seen in Table 3, Columns (2) and (4), although baseline levels of opioid use disorder are very different, the relative risk reduction in the probability of developing opioid use disorder after PDMP introduction across both groups is extremely similar – 0.0635 and 0.0670. In other words, both groups felt the same net benefit from the introduction of the PDMP, in terms of their relative reductions in the probability of developing opioid use disorder. This result means that from the perspective of a policymaker, targeting the >98.5% group for a reduction in prescribing doesn't have any benefit over reducing prescribing for the <98.5% group, or reducing prescribing for the population broadly. To see this, consider a policymaker that can implement an opioid-reduction policy that has causal effects of the same order of magnitude as the quasi-experiment here, but they must choose to do so for only one of the two groups. If they chose to take aim at the >98.5% group, $1,085,907 * -.0025758 = 2,797$ cases of OUD would be averted. And if they took aim at the <98.5% group, $52,434,150 * -.000053 = 2,779$ cases would be averted. The algorithm thus does not seem to identify heterogeneous groupings of pa-

estimated absolute effects here are simply scaled by each group's baseline risk – the seemingly large difference is simply caused by "risk magnification." The RRR calculation of $RRR^1 = -\frac{-0.025}{0.05} = 0.5$ and $RRR^2 = -\frac{-0.002775}{0.00555} = 0.5$ are the correct objects of interest to show that there is no heterogeneity in treatment effects in this setting. To further highlight the misguidedness of using absolute effects for setting policy, imagine a policymaker had to implement a policy to achieve a 50% reduction in prescribing. Suppose they followed the above findings on absolute effects being larger in group 1, and thus decide to take all opioids away from Group 1. 50 people would no longer receive an opioid, and the reduction in opioid use disorder would be 5 individuals, or 50%. But this would be no more impactful than taking away all opioids from Group 2, or from simply taking away opioids from half of all individuals in the full 1,000-person group at random. In order for the policymaker's decision to target specific groups to be sound (and have an effect more impactful than choosing people at random), they need to take away opioids from a group with *high heterogeneous treatment effects* as estimated by a demonstrably large *relative risk reduction*.

tients that most benefit from reducing opioid prescribing, instead, every group appears to get the same broad benefit. If anything, it is the opposite. We calculate an "implied treatment effect" in the last line of Table 3. For this analysis, we assume that reductions in prescribing brought about by PDMP implementation are causally and proportionally creating the observed reductions in opioid use disorder. (This interpretation relies, as in traditional IV frameworks, on an exclusion restriction – that the only channel by which PDMPs influence rates of OUD is via the opioid prescribing channel.) The implied treatment effect is calculated as:

$$ITE^r = \frac{RRR^r_{oud}}{RRR^r_{rx}} \tag{9}$$

or the relative risk reduction for opioid use disorder divided by the relative risk reduction for opioid prescribing for cohort r. This is the implied proportional reduction in OUD rates brought about by a concomitant proportional reduction in prescribing for stratum r. As shown in Table 3, the results of this exercise seem to indicate that a greater "bang for buck" in terms of prescribing may be found in the *non-targeted* (<98.5%) population: they saw the same relative reduction in OUD as the targeted group, but the associated relative reduction in opioid prescribing was smaller.

4.4 Results for 10 strata

Finally, we depict a similar analysis for 10 risk strata instead of two, partitioning the individuals into ten groups according to their machine-predicted risk score. As above, group sizes are calibrated for convenience such that moving up one risk stratum selects approximately 10% more of the true positives into the sample. In Appendix Figure A.3, we further repeat the same exercise for 20 strata to show the conclusions do not change based on the arbitrary choice of the number of bins.

Figure 6 depicts results in three panels. The top panel displays the relative risk reduction in prescribing versus the machine-predicted risk score, and the middle panel the relative risk reduction in opioid use disorder versus the machine-predicted risk score. As can be seen, the machine-predicted risk scores *do* again correlate to doctor decisions: patients targeted by the algorithm see the greatest reduction in opioid prescribing after the introduction of a PDMP, indicating doctors are to some degree in agreement with the algorithm regarding which patients are most at risk.[9] However, the relative risk reduction in the development of OUD is completely uncorrelated to the scores generated by the algorithm, indicating it is doing an extremely poor job at identifying groups with high heterogeneous

[9]Recall that relative risk reduction, RRR, is defined as the negative of the point estimate of the reduction, divided by the pre-period mean. Thus the upward sloping graphs represent increasingly large reductions as the risk score increases.

treatment effects. Fitting a line to the coefficients in the top panel yields a p-value of 0.006, indicating a strong correlation between machine risk scores and doctor behavior in terms of opioid prescribing reductions after the introduction of a PDMP. But the line fit to the coefficients on the middle panel has a p-value of 0.842, indicating no correlation between the machine-predicted risk score and reductions in opioid use disorder, and thus no correlation between predicted risk scores and heterogeneous treatment effects, the object of interest.

Finally, in the bottom panel, we again report "implied treatment effects" as defined above in Equation 9. Once again, as above, if anything, the implied treatment effect seems to be highest in the bin deemed lowest-risk by the model. (The same effect can be observed even more strongly in the 20-bin specification found in Appendix Figure A.3.)

4.5 Discussion

The results above strongly suggest that machine learning models currently being developed to identify individuals at risk of opioid use disorder and inform clinical decision making around the prescribing of opioids are not well-specified to generate individual scores that correlate with the object of interest, an individual's heterogeneous treatment effect of receiving an opioid.

The results indicate that if doctors were to reallocate prescribing according to the algorithm's recommendations – prescribing fewer opioids to chronic pain patients and more to patients identified as "low-risk", they would at best make no improvements on existing decisionmaking, and perhaps actually *worsen* rates of development of opioid use disorder.

Further, in Section 2.2.1, we noted that the targeted populations are largely characterized by the presence of complex chronic pain conditions and other burdensome comorbidities. Chronic pain patients have a higher rate of opioid receipt, and have a higher baseline risk of opioid use disorder, and the algorithm has correctly identified them as such. But as shown above, this targeted group does not appear to experience abnormally high benefits to having prescribing reduced, in terms of future *reductions* in OUD. A reduction in prescribing brought about by the introduction of PDMPs yielded no more benefits in this group than any other.

In this context, a major additional concern is that this targeted group (risk scores >98.5%) are likely to experience the highest *costs* of having prescribing reduced, in terms of untreated pain, as many chronic pain patients on long-term opioid therapy are still on that therapy because they have not found other non-pharmocological alternatives. In the welfare model presented in Section 3.1, these individuals

may have an especially high C_H.

By contrast, opioid consumption in the non-targeted group (<98.5%) occurs at very low rates, and this group likely is comprised of patients who receive opioids incidentally for their acute pain needs, such as wisdom tooth removal. Although the algorithm correctly identifies these individuals as having a very low overall *baseline* risk of OUD, as discussed above, their relative risk reduction is nearly identical to that of the chronic pain patients in the targeted group, and if anything their implied treatment effect is *greater*. Further, they may have a relatively low C_H, making reductions in this group less harmful to social welfare.

As such this group may actually represent a fruitful area for targeting prescribing reductions, where the benefits are equal to or greater than other types of individuals, and costs are relatively low. In line with this observation, many policymakers have identified overprescribing of opioids for acute conditions as a major area of concern; legislative approaches to fix this often center on strict restrictions on the days supply that may be prescribed for an acute condition or after a medical procedure like surgery. While not without potential costs to pain experienced and lost productivity, this may represent an approach with greater overall social welfare than the incorporation of opioid risk models into clinical decision support.

Finally, an important consideration is the dynamic impacts of implementing these models into clinical practice on physician behavior. A physician confronted with an algorithmic recommendation may feel constrained or overriden by the recommendation, and may worry that disregarding or going against the machine recommendation might expose them to legal liability or other legal scrutiny. This effect might present problems in both directions: a patient inappropriately labeled "high risk" might be denied critical pain therapy. Conversely, a patient may be inappropriately labeled "low-risk," and the ML model may perversely increase the clinician's confidence in prescribing an opioid to a person at relatively high risk.

5 Conclusion

In this paper, we trained a machine learning model to predict opioid use disorder that mirrors many currently being deployed to provide treatment decision support to doctors. The social value of these new algorithms will depend on whether the estimated risk scores are correlated with heterogeneous causal treatment effects of receiving an opioid prescription.

We plug our machine-generated opioid risk scores into a quasi-experimental setting to investigate

whether risk scores correlate with heterogeneous treatment effects, finding that they do an extremely poor job of generating a cohort stratification with high versus low heterogeneous treatment effects. Results suggest that if doctors were to reallocate prescribing to become even more in concordance with risk scores, it is unlikely that improvements in opioid use disorder outcomes will be superior than if the reallocation happened randomly. Perversely, rates of opioid use disorder might actually *worsen*.

We add to the literature on many and varied forms of algorithmic unfairness: here, unfairness arises from incorrect choice of the objective function. This form of unfairness may be particularly likely with models to predict opioid use disorder, but may be relevant to many applications across healthcare and Artificial Intelligence, where models for clinical use are developed around machine identification of baseline risk rather than heterogeneous treatment effects.

References

Alford, Daniel P. 2016. "Opioid Prescribing for Chronic Pain – Achieving the Right Balance through Education." *New England Journal of Medicine* 374 (4):301–303.

Appriss Health. 2018. "Statewide Opioid Assessment: Michigan. Identify, Prevent, and Manage Substance Use Disorders."

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics* 47 (2):1148–1178.

Athey, Susan and Stefan Wager. 2019. "Estimating Treatment Effects with Causal Forests: An Application." *arXiv:1902.07409 [stat]* .

Barocas, Joshua A., Laura F. White, Jianing Wang, Alexander Y. Walley, Marc R. LaRochelle, Dana Bernson, Thomas Land, Jake R. Morgan, Jeffrey H. Samet, and Benjamin P. Linas. 2018. "Estimated Prevalence of Opioid Use Disorder in Massachusetts, 2011–2015: A Capture–Recapture Analysis." *American Journal of Public Health* 108 (12):1675–1681.

Brenton, Ashley, Steven Richeimer, Maneesh Sharma, Chee Lee, Svetlana Kantorovich, John Blanchard, and Brian Meshkin. 2017. "Observational Study to Calculate Addictive Risk to Opioids: A Validation Study of a Predictive Algorithm to Evaluate Opioid Use Disorder." *Pharmacogenomics and Personalized Medicine* 10:187–195.

Bruehl, Stephen, A. Vania Apkarian, Jane C. Ballantyne, Ann Berger, David Borsook, Wen G. Chen, John T. Farrar, Jennifer A. Haythornthwaite, Susan D. Horn, Michael J. Iadarola, Charles E. Inturrisi, Lixing Lao, Sean Mackey, Jianren Mao, Andrea Sawczuk, George R. Uhl, James Witter, Clifford J. Woolf, Jon-Kar Zubieta, and Yu Lin. 2013. "Personalized Medicine and Opioid Analgesic Prescribing for Chronic Pain: Opportunities and Challenges." *The Journal of Pain* 14 (2):103–113.

Centers for Disease Control. 2018. "Calculating Total Daily Dose of Opioids For Safer Dosage." Tech. rep.

Cepeda, M. Soledad, Daniel Fife, Wing Chow, Gregory Mastrogiovanni, and Scott C. Henderson. 2012. "Assessing Opioid Shopping Behaviour: A Large Cohort Study from a Medication Dispensing Database in the US." *Drug Safety* 35 (4):325–334.

Chaparro, Luis Enrique, Andrea D Furlan, Amol Deshpande, Angela Mailis-Gagnon, Steven Atlas, and Dennis C Turk. 2014. "Opioids Compared with Placebo or Other Treatments for Chronic Low Back Pain: An Update of the Cochrane Review." *Spine* 39 (7):556–563.

Che, Zhengping, Jennifer St Sauver, Hongfang Liu, and Yan Liu. 2017. "Deep Learning Solutions for Classifying Patients on Opioid Use." *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2017:525–534.

Chou, Roger, Jane C. Ballantyne, Gilbert J. Fanciullo, Perry G. Fine, and Christine Miaskowski. 2009. "Research Gaps on Use of Opioids for Chronic Noncancer Pain: Findings From a Review of the Evidence for an American Pain Society and American Academy of Pain Medicine Clinical Practice Guideline." *The Journal of Pain* 10 (2):147–159.e15.

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

Dueñas, María, Begoña Ojeda, Alejandro Salazar, Juan Antonio Mico, and Inmaculada Failde. 2016. "A Review of Chronic Pain Impact on Patients, Their Social Environment and the Health Care System." *Journal of Pain Research* 9:457–467.

Evans, William N., Ethan M. J. Lieber, and Patrick Power. 2019. "How the Reformulation of OxyContin Ignited the Heroin Epidemic." *The Review of Economics and Statistics* 101 (1):1–15.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2009. *The Elements of Statistical Learning*. Springer, Berlin: Springer Series in Statistics, second ed.

Furlan, Andrea D., Juan A. Sandoval, Angela Mailis-Gagnon, and Eldon Tunks. 2006. "Opioids for Chronic Noncancer Pain: A Meta-Analysis of Effectiveness and Side Effects." *CMAJ: Canadian Medical Association Journal* 174 (11):1589–1594.

Harrell, Frank. 2018. "Viewpoints on Heterogeneity of Treatment Effect and Precision Medicine." Library Catalog: www.fharrell.com.

Hasan, Md Mahmudul, Md Noor-E-Alam, Mehul Rakeshkumar Patel, Alicia Sasser Modestino, Leon D. Sanchez, and Gary Young. 2019. "A Novel Big Data Analytics Framework to Predict the Risk of Opioid Use Disorder." *arXiv:1904.03524 [cs, q-bio, stat]* .

Hastings, Justine S., Mark Howison, and Sarah E. Inman. 2020. "Predicting High-Risk Opioid Prescriptions before They Are Given." *Proceedings of the National Academy of Sciences* 117 (4):1917–1923.

Hedegaard, Holly, Arialdi M. Miniño, and Margaret Warner. 2020. "Drug Overdose Deaths in the United States, 1999–2018." Tech. Rep. 356, National Center for Health Statistics, Hyattsville, MD.

Hudson, Sally, Peter Hull, and Jack Liebersohn. 2017. "Interpreting Instrumented Difference-in-Differences."

Katz, Nathaniel, Lee Panas, MeeLee Kim, Adele D. Audet, Arnold Bilansky, John Eadie, Peter Kreiner, Florence C Paillard, Cindy Thomas, and Grant Carrow. 2010. "Usefulness of Prescription Monitoring Programs for Surveillance-Analysis of Schedule II Opioid Prescription Data in Massachusetts, 1996-2006." *Pharmacoepidemiology and Drug Safety* 19 (2):115–123.

Kilby, Angela. 2019. *Opioids for the Masses: Welfare Tradeoffs in the Regulation of Narcotic Pain Medications*. Ph.D. thesis, Massachusetts Institute of Technology.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105 (5):491–495.

Lada, Akos, Alexander Peysakhovich, Diego Aparicio, and Michael Bailey. 2019. "Observational Data for Heterogeneous Treatment Effects with Application to Recommender Systems." In *Proceedings of the 2019 ACM Conference on Economics and Computation - EC '19*. Larnaca, Cyprus: ACM Press, 199–213.

Lin, Dora H., Eleanor Lucas, Irene B. Murimi, Katherine Jackson, Michael Baier, Shannon Frattaroli, Andrea C. Gielen, Patience Moyo, Linda Simoni-Wastila, and G. Caleb Alexander. 2017. "Physician Attitudes and Experiences with Maryland's Prescription Drug Monitoring Program (PDMP): Physician Experiences with PDMPs." *Addiction* 112 (2):311–319.

Lo-Ciganic, Wei-Hsuan, 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. 2019. "Evaluation of Machine-Learning Algorithms for Predicting Opioid Overdose Risk Among Medicare Beneficiaries With Opioid Prescriptions." *JAMA Network Open* 2 (3):e190968.

Morgan, Daniel J., Bill Bame, Paul Zimand, Patrick Dooley, Kerri A. Thom, Anthony D. Harris, Soren Bentzen, Walt Ettinger, Stacy D. Garrett-Ray, J. Kathleen Tracy, and Yuanyuan Liang. 2019. "As-

sessment of Machine Learning vs Standard Prediction Rules for Predicting Hospital Readmissions." *JAMA Network Open* 2 (3):e190348.

Nicholson, Kate M., Diane E. Hoffman, and Chad D. Kollas. 2018. "How the CDC's Opioid Prescribing Guideline Is Harming Pain Patients."

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464):447–453.

Oh, Jeeheh, Maggie Makar, Christopher Fusco, Robert McCaffrey, Krishna Rao, Erin E. Ryan, Laraine Washer, Lauren R. West, Vincent B. Young, John Guttag, David C. Hooper, Erica S. Shenoy, and Jenna Wiens. 2018. "A Generalizable, Data-Driven Approach to Predict Daily Risk of *Clostridium Difficile* Infection at Two Large Academic Health Centers." *Infection Control & Hospital Epidemiology* 39 (4):425–433.

Ohio Board of Pharmacy. 2019. "Ohio PDMP AWARE User Support Manual."

Parente, Stephen T., Susan S. Kim, Michael D. Finch, Lisa A. Schloff, Thomas S. Rector, Raafat Seifeldin, and J. David Haddox. 2004. "Identifying Controlled Substance Patterns of Utilization Requiring Evaluation Using Administrative Claims Data." *The American Journal of Managed Care* 10 (11 Pt 1):783–790.

Powers, Scott, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H. Shah, Trevor Hastie, and Robert Tibshirani. 2018. "Some Methods for Heterogeneous Treatment Effect Estimation in High Dimensions: Some Methods for Heterogeneous Treatment Effect Estimation in High Dimensions." *Statistics in Medicine* 37 (11):1767–1787.

Ravindranath, Mohana. 2019. "How Your Health Information Is Sold and Turned into 'Risk Scores'." *Politico* .

Rough, Kathryn, Krista F. Huybrechts, Sonia Hernandez-Diaz, Rishi J. Desai, Elisabetta Patorno, and Brian T. Bateman. 2019. "Using Prescription Claims to Detect Aberrant Behaviors with Opioids: Comparison and Validation of 5 Algorithms." *Pharmacoepidemiology and Drug Safety* 28 (1):62–69.

Ruhm, Christopher J. 2018. "Corrected US Opioid-Involved Drug Poisoning Deaths and Mortality Rates, 1999-2015: Corrected Opioid-Involved Mortality Rates." *Addiction* 113 (7):1339–1344.

Speights, David and Ray Atencio. 2018. "Applying Machine Learning and Analytics to Combat the Opioid Crisis." Tech. rep., Appriss Health.

Sullivan, Mark D., Mark J. Edlund, Ming-Yu Fan, Andrea Devries, Jennifer Brennan Braden, and Bradley C. Martin. 2010. "Risks for Possible and Probable Opioid Misuse among Recipients of Chronic Opioid Therapy in Commercial and Medicaid Insurance Plans: The TROUP Study." *Pain* 150 (2):332–339.

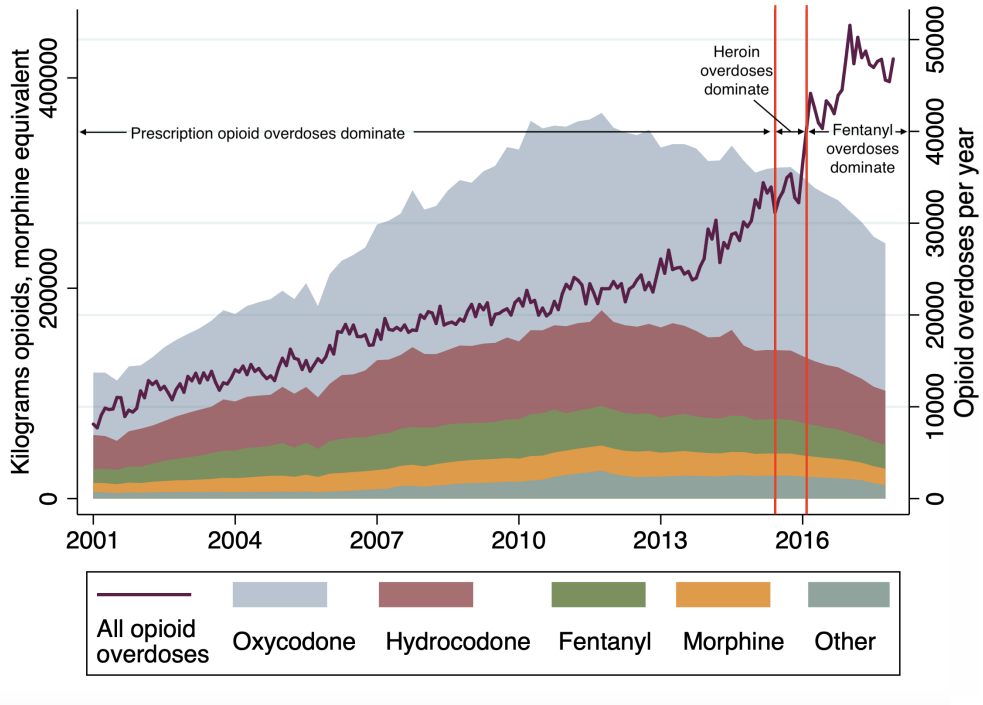
U.S. Department of Justice. 2015. "Justice System Use of Prescription Drug Monitoring Programs: Call To Action and Issue Brief." Tech. rep.

Wager, Stefan and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523):1228–1242.

Wiens, Jenna, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N. Ossorio, Sonoo Thadaney-Israni, and Anna Goldenberg. 2019. "Do No Harm: A Roadmap for Responsible Machine Learning for Health Care." *Nature Medicine* 25 (9):1337–1340.

Figures

Figure 1: Opioids dispensed via pharmacies, hospitals, practitioners, etc., overlayed with opioid overdoses, from 2001-2017



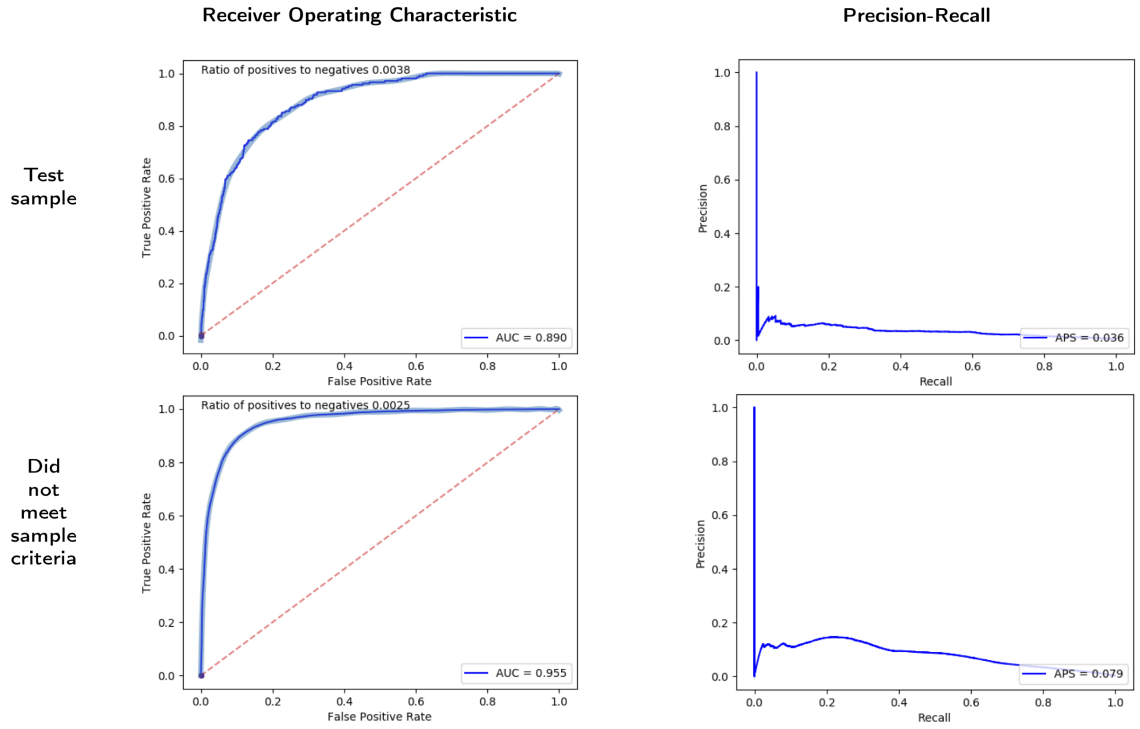
Notes: Opioid dispensing data is compiled from the DEA's ARCOS reporting system, and depicts total kilograms of opioids dispensed in morphine equivalents, and the proportion of that total accounted for by oxycodone, hydrocodone, etc. Each opioid is converted to morphine equivalents using a conversion factor, morphine milligram equivalents, or MME (Centers for Disease Control, 2018). Opioid overdose deaths are from CDC's NCHS Vital Statistics mortality data. Opioid overdose deaths are deaths with underlying cause of death code of X40-X44, X60-X64, X85, or Y10-Y14 (indicating a drug overdose), and multiple causes of death codes of T40.0-T40.4 or T40.6 (indicating opioids or heroin specifically were listed). "Prescription opioid overdoses dominate" from 2001 to mid-2015, when T40.2 – semisynthetic opioids, including drugs such as oxycodone, hydrocodone, hydromorphone, and oxymorphone – was the most frequently mentioned drug on a death certificate. "Heroin overdoses dominate" from mid-2015 to early 2016 when heroin (T40.1) was most frequently mentioned, and "fentanyl overdoses dominate" when T40.4 – synthetic opioid analgesics other than methadone, including drugs such as fentanyl and tramadol – is most frequently listed. Note that many overdoses are poly-drug overdoses and therefore include more than one T-code.

Figure 2: Feature importances from gradient boosted model

Total Gain	Variable description	Total Gain	Variable description
24.93	Age 26-31	7.81	Industry: Finance, Insurance, Real Estate
24.84	Age 31-36	7.40	Spondylosis and allied disorders
24.64	Chiropractor/DCM	7.21	Migraine
24.37	Industry: Manufacturing, Nondurable Goods	6.98	Provider Therapy (Physical)
23.80	Organic Sleep Disorders	6.93	Drug dependence
21.82	Industry: Transportation, Communications, Utilities	6.06	Anxiety, dissociative and somatoform disorders
20.84	Provider: Family Practice	5.65	Provider: Pain Mgmt/Pain Medicine
20.79	Intervertebral disc disorders	5.39	Place of service: Other Unlisted Facility
20.09	Industry: Services	5.27	Provider: Home Health Organiz/Agency
18.83	Sex: Male	5.26	Provder: Rheumatology
18.60	Other and unspecified disorders of back	5.02	Provider: Physical Medicine & Rehab
18.52	Age 21-26	4.99	Sarcoidosis
15.16	Nonallopathic lesions, not elsewhere classified	4.85	Age 16-21
14.67	Age 46-51	4.76	Place of service: Psych Facility Partial Hosp
13.92	Place of service: Outpatient Hospital	4.60	Testicular dysfunction
13.22	Industry: Manufacturing, Durable Goods	4.60	Other congenital anomalies of circulatory system
12.93	Age 41-46	4.44	Other and unspecified infectious and parasitic diseases
11.84	Provider: Medical Doctor - MD (NEC)	4.30	Provider: Neurology
11.75	Viral hepatitis	4.22	Place of service: Comprehensive Outpt Rehab Fac
11.63	Diseases of nail	4.19	Sprains and strains of other and unspecified parts of back
11.56	Age 51-56	4.10	General symptoms
10.01	Age 56-61	4.08	Malignant neoplasm of liver and intrahepatic bile ducts
10.00	Sprains and strains of ankle and foot	3.93	Provider: Psychologist
9.99	Industry: Retail Trade	3.88	Other extrapyramidal disease and abnormal movement disorders
9.47	Other disorders of cervical region	3.85	Industry: Oil & Gas Extraction, Mining
9.29	Provider: Anesthesiology	3.67	Place of service: Military Treatment Facility
8.71	Provider: Osteopathic Medicine	3.44	Provider: Otolaryngology
8.39	Age 36-41	3.41	Encounter for other and unspecified procedures and aftercare
8.37	Malignant neoplasm of floor of mouth	3.29	Other diseases of blood and blood-forming organs
7.83	Place of service: Patient Home	3.29	Artificial opening status

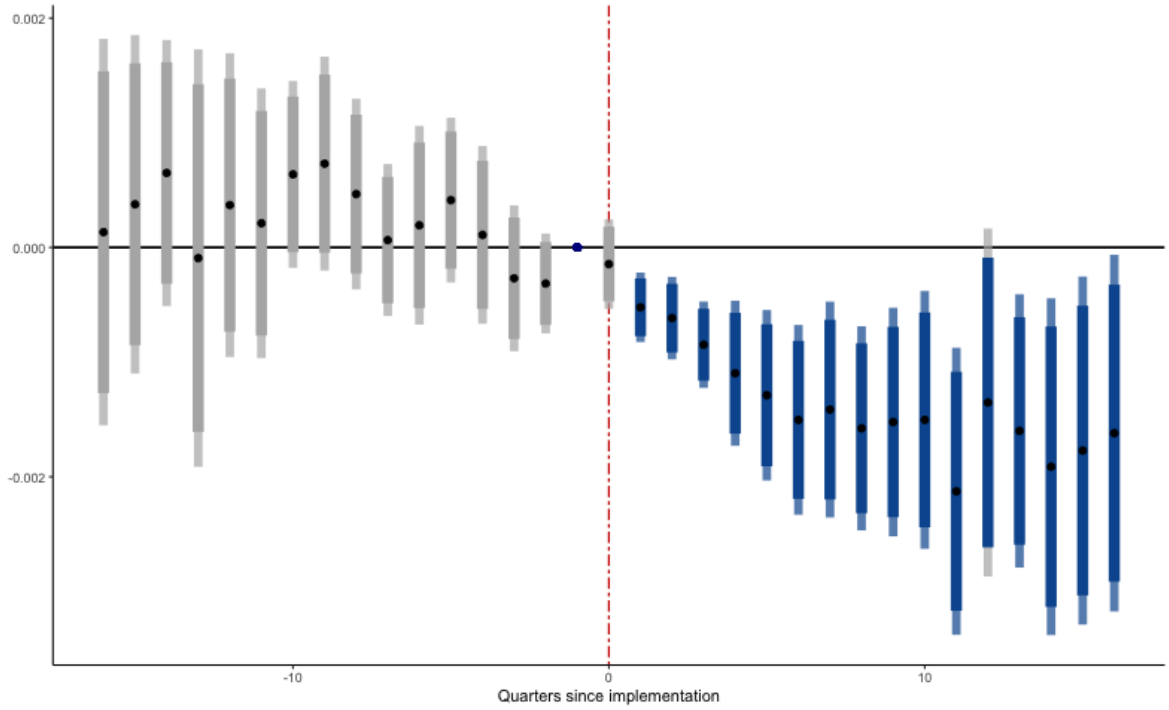
Notes: Feature importance measured according to the "total gain," the average training loss reduction achieved by the feature when used for splitting.

Figure 3: Performance metrics of the machine learning algorithm – receiver operating characteristic (ROC) and precision-recall curves, constructed at the individual level



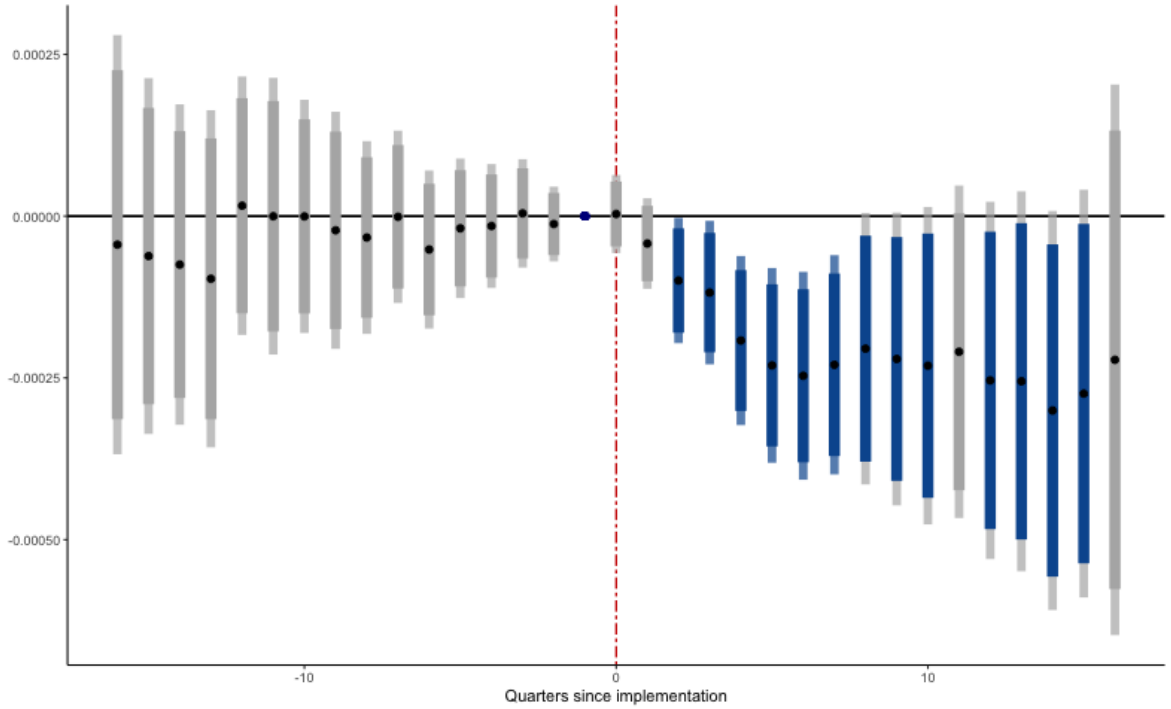
Notes: The left panel depicts receiver operating characteristic curves for the 10% hold-out test sample (top panel) and the enrollees who did not meet the cohort criteria (bottom panel). The receiver operating characteristic curve represents the tradeoff between a high true positive rate and a low false positive rate, for the range of cutoff thresholds that could be chosen by the researcher. The right panel depicts the precision-recall curve, which represents the tradeoff between precision, or positive predictive power (the percent of enrollees labeled as positive who are actually positive), and recall (sensitivity, or the percent of positive individuals correctly labeled as such.)

Figure 4: Dynamic treatment effect of the impact of PDMP introduction on opioid prescribing for the full sample



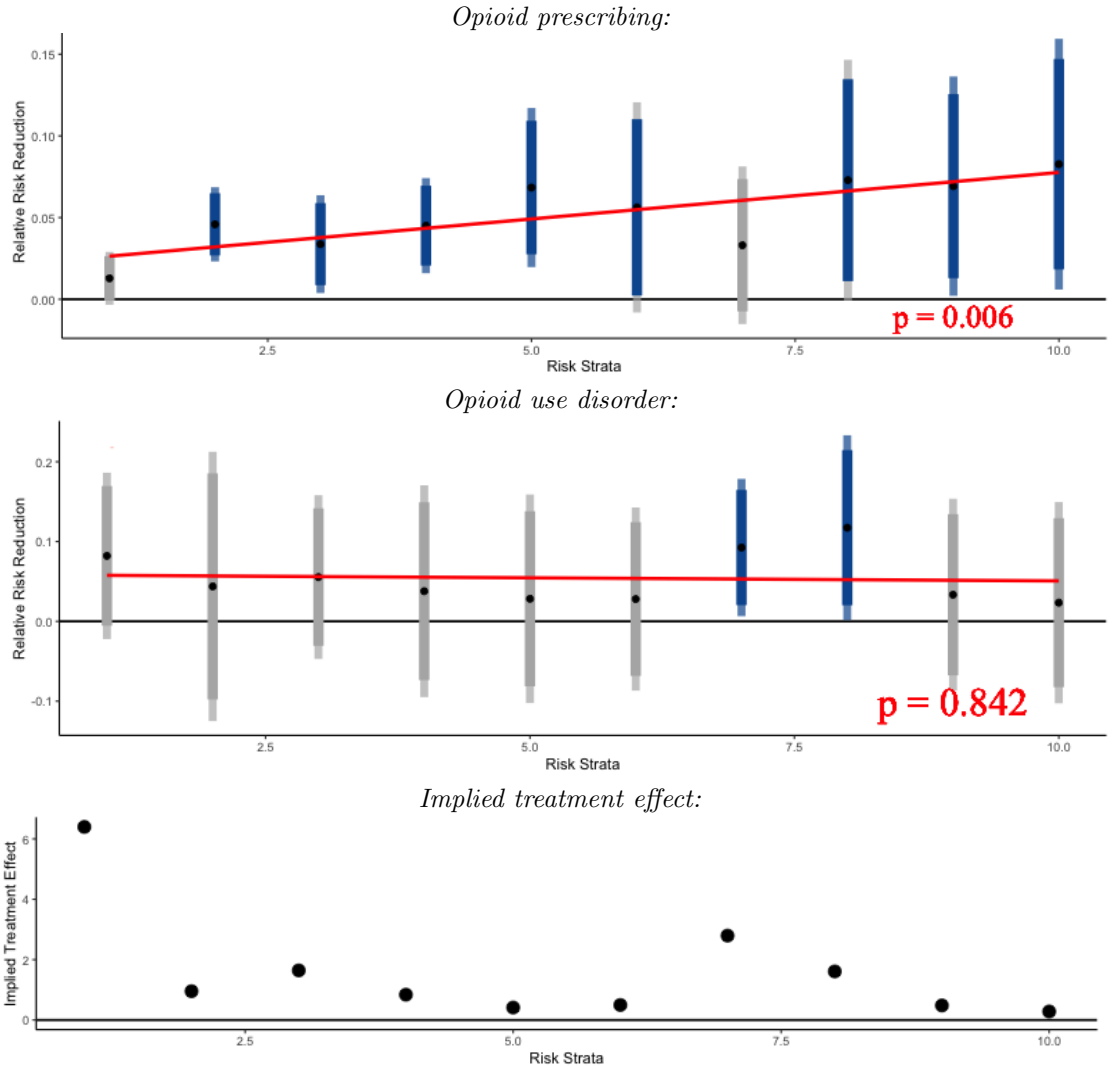
Notes: The figure depicts an event study where leads and lags of the impact of PDMP introduction on opioid prescribing are estimated according to Equation 7 in the full sample of all insured enrollees (one risk stratum). Bars represent 90% and and 95% confidence intervals, and robust standard errors are clustered at the state level.

Figure 5: Dynamic treatment effect of the impact of PDMP introduction on opioid use disorder for the full sample



Notes: The figure depicts an event study where leads and lags of the impact of PDMP introduction on opioid use disorder are estimated according to Equation 7 in the full sample of all insured enrollees (one risk stratum). Bars represent 90% and and 95% confidence intervals, and robust standard errors are clustered at the state level.

Figure 6: Relative risk reduction across 10 strata



Notes: The figure depicts point estimates from ten separate regressions estimating the impact of PDMP introduction on outcomes according to Equation 8, for risk strata $r \in [1, 10]$. Relative risk reduction $RRR \approx -\frac{\text{Estimated effect}}{\text{Pre-Mean}}$ and the approximate relative risk reduction is calculated in Stata using postestimation of marginal effects (`margins`) and a non-linear combination of estimators (`nlcom`) to report the mean and standard error of the main effect divided by the pre-period mean. The top panel contains RRR by risk score strata for opioid prescribing, and the middle panel for opioid use disorder. The bottom panel contains an estimate for the implied treatment effect in each risk strata, or the estimate of RRR for opioid use disorder divided by the estimate of RRR for opioid prescribing brought about by the introduction of a PDMP.

Tables

Table 1: The impact of PDMP introduction on opioid prescribing for the full sample

	Received Opioids
Post-PDMP	-0.000874
	(0.000210)
Controls:	
Ind. Controls	Y
Ind. FE	N
Entry-cohort FE	Y
Quarter FE	Y
Enrollees	4,944,277
States	38
Observations	53,520,057
Pre-period Mean	0.0258
−(Effect/Pre-Mean)	
≈ Relative	0.0349
Risk Reduction	(0.0083)

Notes: The table reports the impact of PDMP introduction on opioid prescribing, and is estimated according to Equation 8 in the full sample of all insured enrollees (one risk stratum). The approximated relative risk reduction is calculated in Stata using postestimation of marginal effects (`margins`) and a non-linear combination of estimators (`nlcom`) to report the mean and standard error of the main effect divided by the pre-period mean. Robust standard errors, clustered at the state level, are in parentheses.

Table 2: The impact of PDMP introduction on opioid use disorder for the full sample

	Opioid Use Disorder claim
Post-PDMP	-0.000140
	(0.000070)
Controls:	
Ind. Controls	Y
Ind. FE	N
Entry-cohort FE	Y
Quarter FE	Y
Enrollees	4,944,277
States	38
Observations	53,520,057
Pre-period Mean	0.0015
−(Effect/Pre-Mean)	
≈ Relative	0.0871
Risk Reduction	(0.0420)

Notes: The table reports the impact of PDMP introduction on opioid use disorder, and is estimated according to Equation 8 in the full sample of all insured enrollees (one risk stratum). The approximated relative risk reduction is calculated in Stata using postestimation of marginal effects (`margins`) and a non-linear combination of estimators (`nlcom`) to report the mean and standard error of the main effect divided by the pre-period mean. Robust standard errors, clustered at the state level, are in parentheses.

Table 3: Impact of a reduction in opioid prescribing on two risk strata

	Non-targeted group		Targeted group		Full Sample	
	(1) Received Opioids	(2) Opioid Use Disorder claim	(3) Received Opioids	(4) Opioid Use Disorder claim	(5) Received Opioids	(6) Opioid Use Disorder claim
Post-PDMP	-0.000716	-0.000053	-0.005656	-0.002578	-0.000874	-0.000140
	(0.000179)	(0.000044)	(0.001875)	(0.001169)	(0.000210)	(0.000070)
Enrollees	4,871,699	4,871,699	72,578	72,578	4,944,277	4,944,277
States	38	38	38	38	38	38
Observations	52,434,150	52,434,150	1,085,907	1,085,907	53,520,057	53,520,057
Pre-period Mean	0.0242	0.0008	0.0976	0.0357	0.0258	0.0015
−(Effect/Pre-Mean) ≈ Relative Risk Reduction	0.0304	0.0635	0.0587	0.0670	0.0349	0.0871
	(0.0076)	(0.0513)	(0.0190)	(0.0296)	(0.0083)	(0.0420)
Implied Treatment Effect	2.0866		1.1413		2.4948	

Notes: The table depicts results from three separate regressions estimating the impact of PDMP introduction on outcomes according to Equation 8, for risk strata r above and below the 98.5 percentile of all risk scores, as well as the full sample. The "non-targeted group" is below the 98.5 percentile of all scores, and the "targeted group" is above; each group represents about half of all OUD cases, i.e. the cutoff is set so that the sensitivity of the model is ≈50%. The approximate relative risk reduction is calculated in Stata using postestimation of marginal effects (`margins`) and a non-linear combination of estimators (`nlcom`) to report the mean and standard error of the main effect divided by the pre-period mean. The bottom row contains an estimate for the implied treatment effect in each risk strata, or the estimate of RRR for opioid use disorder divided by the estimate of RRR for opioid prescribing brought about by the introduction of a PDMP. Robust standard errors, clustered at the state level, in parentheses.

A Appendix

Figure A.1: Truven Marketscan variables that indicate a claim was in a mental health or substance abuse facility

Admission Type — Type of hospital admission.

4: Psych & Substance Abuse

Service Sub-Category Code — A code indicating a detailed category of service

31110	Substance Abuse Facility IP Room and Board
31115	Substance Abuse Facility IP Procedures
31118	Substance Abuse Facility IP Behavioral Health Therapy
31120	Substance Abuse Facility IP ER
31130	Substance Abuse Facility IP Diagnostic Services
31764	Substance Abuse OP Nuclear Medicine
31765	Substance Abuse OP PET Scans
31766	Substance Abuse OP Therapeutic Radiology
31767	Substance Abuse OP Ultrasounds
31768	Substance Abuse OP X-Rays
31769	Substance Abuse OP Radiology Other
30110	Mental Health Facility IP Room and Board
30115	Mental Health Facility IP Procedures
30118	Mental Health Facility IP Behavioral Health Therapy
30120	Mental Health Facility IP ER
30130	Mental Health Facility IP Diagnostic Services
30131	Mental Health Facility IP Dialysis
30132	Mental Health Facility IP DME

Provider Type

20	Mental Health/Chemical Dep NEC
21	Mental Health Facilities
22	Chemical Depend Treatment Ctr
23	Mental Hlth/Chem Dep Day Care
25	Rehabilitation Facilities
35	Residential Treatment Center

Place of Service — Setting where service occurred.

51	Inpatient Psychiatric Facility
52	Psych Facility Partial Hosp
53	Community Mental Health Center
54	Intermed Care/Mental Retarded
55	Residential Subst Abuse Facil
56	Psych Residential Treatmnt Ctr
57	Non-resident Subst Abuse Facil
61	Comprehensive Inpt Rehab Fac
62	Comprehensive Outpt Rehab Fac

Figure A.2: 38 states which implemented PDMP after 2002 or have not implemented - excludes early controlled substances monitoring efforts

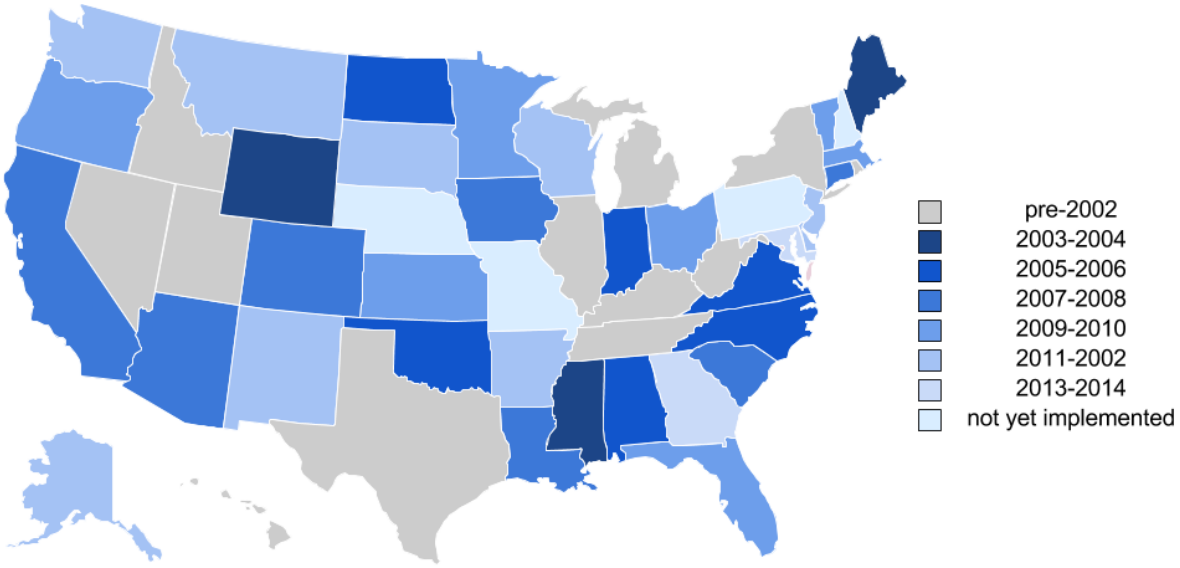
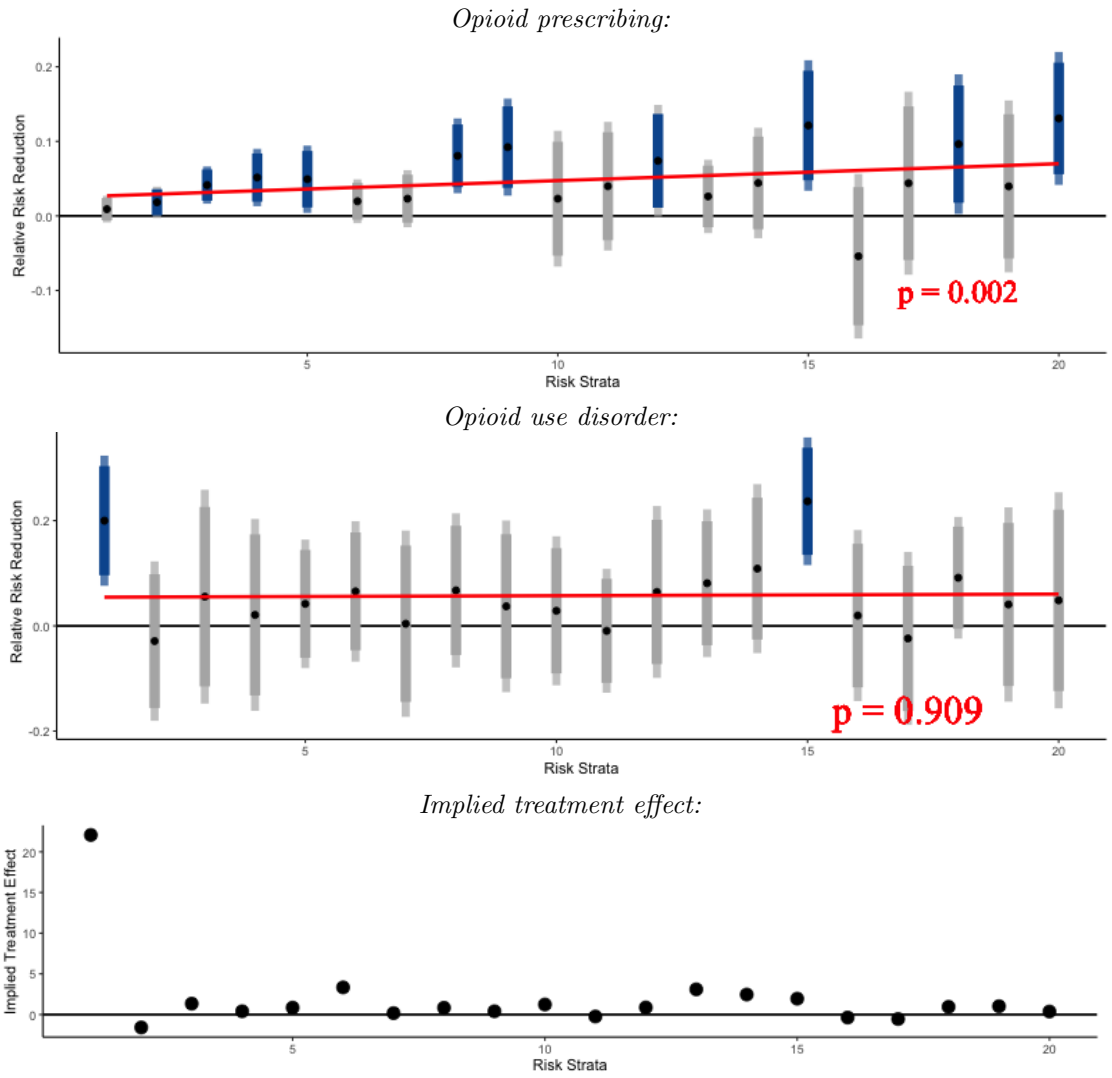


Figure A.3: Relative risk reduction across 20 strata



Notes: The figure depicts point estimates from twenty separate regressions estimating the impact of PDMP introduction on outcomes according to Equation 8, for risk strata $r \in [1, 20]$. Relative risk reduction $RRR \approx -\frac{\text{Estimated effect}}{\text{Pre-Mean}}$ and the approximate relative risk reduction is calculated in Stata using postestimation of marginal effects (`margins`) and a non-linear combination of estimators (`nlcom`) to report the mean and standard error of the main effect divided by the pre-period mean. The top panel contains RRR by risk score strata for opioid prescribing, and the middle panel for opioid use disorder. The bottom panel contains an estimate for the implied treatment effect in each risk strata, or the estimate of RRR for opioid use disorder divided by the estimate of RRR for opioid prescribing brought about by the introduction of a PDMP.