BMJ Open

# Safe opioid prescribing: a machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada



| | |
|---|---|
| Journal: | *BMJ Open* |
| Manuscript ID | bmjopen-2020-043964 |
| Article Type: | Original research |
| Date Submitted by the Author: | 18-Aug-2020 |
| Complete List of Authors: | Sharma, Vishal; University of Alberta, School of Public Health<br>Kulkarni, Vinaykumar; OKAKI Health Analytics<br>Eurich, Dean; University of Alberta, School of Public Health<br>Kumar, Luke; Alberta Machine Intelligence Institute<br>Samanani, Salim; Okaki Health Intelligence, |
| Keywords: | PUBLIC HEALTH, EPIDEMIOLOGY, Adverse events < THERAPEUTICS, Health & safety < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Clinical governance < HEALTH SERVICES ADMINISTRATION & MANAGEMENT |
| | |

*I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.*

*The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.*

*Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.*

**Safe opioid prescribing: a machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada**

**Author list (in order):**

Vishal Sharma (0000-0001-7907-1183), Vinaykumar Kulkarni, Dean T. Eurich (0000-0003-2197-0463), Luke Kumar, Salim Samanani (0000-0001-6751-4805)

**Address for each author:**

2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of Alberta, Edmonton, Alberta, Canada, T6G 2E1 Vishal Sharma BPharm PhD Candidate,

OKAKI Health Intelligence, Edmonton, Alberta, Canada, Vinaykumar Kulkarni MSc

2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of Alberta, Edmonton, Alberta, Canada, T6G 2E1 Dean Eurich professor

Alberta Machine Intelligence Institute, Edmonton, Alberta, Canada, T5J 3B1 Luke Kumar MSc

OKAKI Health Intelligence, Calgary, Alberta, Canada, Salim Samanani MD, Medical Director

**Corresponding Author:**

Dean Eurich, 2-040 Li Ka Shing Center for Health Research Innovation, University of Alberta, Edmonton, Alberta, Canada, T6G 2E1; Phone 780-492-6333; fax 780-492-7455; email: deurich@ualberta.ca

**Acknowledgement**

This study is based on data provided by The Alberta Strategy for Patient Orientated Research (AbSPORU) SUPPORT unit and Alberta Health. The interpretation and conclusions contained herein are those of the researchers and do not necessarily represent the views of the Government of Alberta or AbSPOR. Neither the Government of Alberta, ABSPOR nor Alberta Health expresses any opinion in relation to this study. This work was supported by Mitacs through the Mitacs Accelerate Program (VS and DTE).

**Contributors:** VS VK LK SS and DTE were involved in the conception and design of the study. VS VK LK SS and DTE analyzed the data. VS VK and LK drafted the article. VS VK LK DTE and SS revised the article. All authors gave final approval of the version to be published. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. DTE is the guarantor.

**Funding:** This study received no funding.



***Competing Interest:*** *All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; Salim Samanani has received grants from the College of Physicians & Surgeons of Alberta; no other relationships or activities that could appear to have influenced the submitted work.*

**Ethical approval:** This study was approved by the Health Research Ethics board at the University of Alberta (#Pro00083807_AME2).

**Data Sharing:** The data used in this study is not available for external analysis. However, administrative health data can be accessed from Alberta Health by following defined research protocols and confidentiality agreements.

**Transparency:** The lead author, VS, (the manuscript's guarantor, Dean Eurich) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as originally planned (and, if relevant, registered) have been explained.

**Word Count: 2659**

## Abstract

**Objective:** To develop machine-learning models employing administrative-health data that can estimate risk of adverse outcomes within 30-days of an opioid dispensation for use by health departments or prescription monitoring programs.

**Design, Setting, and Participants:** This prognostic study was conducted in Alberta, Canada between 2017-2018. Participants included all patients over 18 years of age who received at least one opioid dispensation. Pregnant and cancer patients were excluded.

**Exposure:** Each opioid dispensation served as an exposure.

**Main Outcomes/Measures:** Opioid related adverse outcomes were identified from linked administrative health-data. Machine-learning algorithms were trained using 2017 data to predict risk of hospitalization, emergency department visit, and mortality within 30-days of an opioid dispensation. Two independent validation sets, using 2017 and 2018 data, were used to evaluate model performance. Model discrimination and calibration performance were assessed for all patients and those at higher risk. Machine-learning discrimination was compared to current opioid guidelines.

**Results:** Participants in the 2017 training set (n=275,150) and validation set (n=117,829) had similar baseline characteristics. In the 2017 validation set, c-statistics for the XGBoost, logistic regression, and neural network classifiers were 0.87, 0.87, and 0.80, respectively. In the 2018 validation set (n=393,023), the corresponding c-statistics were 0.88, 0.88, and 0.82. C-statistics from the Canadian guidelines ranged from 0.54-0.69 while the US guidelines ranged from 0.50-0.62. The top 5-percentile of predicted risk for the XGBoost and logistic regression classifiers captured 42% of all events and translated into a post-test probability of 13%, up from the pre-test probability of 1.6%.

**Conclusion:** Machine-learning classifiers, especially incorporating hospitalization/physician claims data, have better predictive performance compared to guideline or prescription history only approaches when predicting 30-day risk of adverse outcomes. Prescription monitoring programs and health departments with access to administrative data can use machine-learning classifiers to effectively identify those at higher risk compared to current guideline-based approaches.

## Article Summary

### Strengths and Limitations:

- This study incorporated near complete capture of opioid dispensations from community pharmacies and used validated administrative health data.
- The study population is the entire provincial population and is generalizable to other populations in Canada and beyond.
- This study used commonly available algorithms to train machine-learning models using data which is available to government health departments in all provinces in Canada and other single payer jurisdictions.
- Our predictive models used dispense events and not medication utilization, which is difficult to capture in administrative data.
- Our training dataset does not account for non-prescription opioids, opioids administered in hospitals, and other risks associated with non-prescription use.

## Introduction

Canada has among the highest rates of opioid prescribing in the world, making prescription opioid use a key driver of the current opioid crisis[1]; a major part of the policy response to the opioid crisis focuses on endorsing safe, appropriate opioid prescribing[2-4]. In order to minimize high risk opioid prescribing and to identify patients at high risk of opioid related adverse outcomes, numerous health regulatory bodies have released clinical practice recommendations for health providers regarding appropriate opioid prescribing[3,5,6].

Prescription monitoring programs (PMPs) have been implemented around the world, like Alberta's provincial Triplicate Prescription Program (TPP)[7] in Canada, and are mandated to monitor the utilization and appropriate use of opioids to reduce adverse outcomes. In most jurisdictions, both population-level monitoring metrics and clinical decision aids are used to identify patients at risk of hospitalization or death and are most often based on prescribing guidelines. However, a comprehensive infrastructure of administrative data containing patient level ICD[8] codes and prescription drug histories exists in Alberta and other provinces in Canada which could be further integrated to predict opioid-related risk. Furthermore, current guidelines' of high risk prescribing and utilization of opioids were derived from studies that used traditional statistical methods (regression analyses) to identify population level risk factors for overdose rather than an individual's absolute risk[3,9,10]; these population estimates may not be generalizable to different populations[11]. Thus, a functional gap exists in many health jurisdictions where much of the available administrative health data is not being leveraged for opioid prescription monitoring.

Supervised machine learning (ML)[12,13] is an approach that uses computer algorithms to build predictive models in the clinical setting that can make use of the large amounts of available administrative data[14,15], all within a well-defined process[16]. Supervised ML trains on labelled data to develop prediction models that are specific to different populations and, in many cases, can provide better predictive performance than traditional, population-based statistical models[10,15,17]. We identified one study[10] that applied ML techniques to predict overdose risk in opioid patients pursuant to a prescription. In their validation sample, they found that the DNN (deep neural network) and GBM (gradient boosting machine) algorithms carried the best discrimination performance based on estimated c-statistics and that the ML approach out-performed the guideline approach in terms of predictive performance.

The objective of our study was to develop and validate ML algorithms to predict the 30-day risk of hospitalization, emergency visit and mortality for a patient in Alberta, Canada at the time of an opioid dispensation using administrative data routinely available to health departments and PMPs. We hypothesized that the ML process would perform better than the current guideline approach for predicting risk of adverse outcomes related to opioid prescribing.

## Methods

### Study Design and Participants

This prognostic study used a supervised ML scheme. All patients in Alberta, Canada who received a dispensation for an opioid, were 18 years of age and older between Jan 1, 2017 and Dec 31, 2018 were eligible. Patients were excluded if they had any previous diagnosis of

cancer, received palliative interventions or were pregnant during the study period (eTable 1 in Supplement) as use of opioids in these contexts is clinically different.

Government health departments and payers in many jurisdictions have systems to capture prescription histories and ICD diagnostic codes. As such, we linked various administrative health data sets available in Alberta, Canada using unique patient identifiers in order to establish a complete description of patient demographics, drug exposures and health outcomes. These databases include *1) Pharmaceutical Information Network (PIN):* PIN data includes all dispensing records from community pharmacies from all prescriber types occurring in the province outside of the hospital setting. PIN collects all drug dispensations irrespective of age or insurance status in Alberta, 2*) Population and Vital Statistics Data (VS, Alberta Services)*: sex, age, date of birth, death date, immigration and emigration data, and underlying cause of death according to the World Health Organization algorithm using ICD codes[8], *3) Hospitalizations and Emergency Department Visits (NACRS [National Ambulatory Care Reporting System], DAD [Discharge Abstract Database]):* all services, length of stay, diagnosis (up to 25 ICD-10[8] based diagnoses). Data and coding accuracy are routinely validated both provincially and centrally via the Canadian Institute for Health Information, and 4*) Physician Visits/Claims (Alberta Health):* date of service, ICD code associated with the claim, procedure and billing information.

This study followed the TRIPOD and STARD reporting guidelines[18-20] and received ethics approval from the University of Alberta ethics board (Pro00083807_AME2). All analyses were done using Python, version 3.7 (Python Software Foundation).

**Outcome**

The primary outcome was a composite of a drug-related hospitalization, emergency department (ED) visit or mortality within 30 days of an opioid dispensation based on ICD-10 codes (T40, F55, F10-19; eTable 2 in Supplement)[2,10].

**Predictor Candidates for ML Models**

Predictor variables in our ML models included those that were informed by the literature[3,4,10] and those directly obtained from the data sets. These included features based on demographics (age, sex, income using Forward Sortation index from postal codes[21]), co-morbidity history using ICD-based Elixhauser score categories[22], health care utilization (number of unique opioid prescribers, number of hospital visits), and drug utilization (level 3 ATC codes[23], oral morphine equivalents[24], concurrent use with benzodiazepines, number of opioid and benzodiazepine dispensations, number of unique opioid and benzodiazepine molecules). Depending on the potential predictor, we used data from 30 days to 5 years before the opioid dispensation to generate model features (eFigure 1 in Supplement). Experiments were performed to identify the features and data sets that contributed most to predicting the outcomes, with a view to minimizing the potential future data requirements for health departments and PMPs.

**Statistical Analyses and Machine-Learning Prediction Evaluation**

We randomly divided the patients in the 2017 portion of our study cohort into training (70%) and validation (30%) sets[13]. Baseline characteristics and event rates were compared in the training vs validation group, and between those who experienced the outcome and those who

did not, using chi-squared tests of independence. As well, we used all 2018 data as another independent validation set.

First, we trained commonly used[13] ML algorithms (eAppendix in Supplement) and tuned model hyperparameters using k-fold (k=5) cross validation to address model overfitting[13,25]. As is common in ML validation studies[10], we reported model discrimination performance using area under the receiver operating characteristic curve (AUROC; c-statistic), positive predictive value (PPV), positive likelihood ratios (PLR), number needed to screen (NNS) and plotted AUROC and precision-recall curves (PRCs). For the more interpretable XGBoost and logistic regression classifiers, we reported feature importance[26] and plotted PRCs that compared all dispenses to those within the top 10 percentiles of estimated risk. As well, for the XGBoost classifier, we described feature impact on model outcome using SHAP values[27,28] to add an additional layer of interpretability. Calibration is crucial in the process of developing a risk predictor[29] so we assessed calibration performance on the 2018 data by dividing the study cohort into percentile categories according to the predicted risk of a dispensation, as was done in previous studies[10,30]. Using the XGBoost and logistic regression classifiers, we analyzed the top 0.1, 1, 5, and 10 percentiles of predicted risk by the number of true and false positives, positive likelihood ratios, post-test probabilities, and number needed to screen. We also performed a simulation of daily data uploads for 2018 Quarter 1 to view the predictive power if a ML risk predictor were to be deployed into a monitoring workflow.

We then compared ML risk prediction to current guideline approaches as others have[10], using the 2019 Centers for Medicare & Medicaid Services opioid safety measures[31] and the 2017 Canadian Opioid Prescribing Guideline[3]. As well, we compared the discrimination

performance of different logistic regression classifier models using various combinations of features derived from their respective databases: **1)** demographic and drug/health utilization features from PIN and 2**)** co-morbidity features derived from DAD, NACRS and Claims.

### Patient and Public Involvement

This research was done without patient involvement. Patients were not invited to comment on the study design and were not consulted to develop patient relevant outcomes or interpret the results. Patients were not invited to contribute to the writing or editing of this document for readability or accuracy. There are no plans to disseminate the results of the research to study participants.

## Results

### Patient Characteristics and Predictors

We identified 392,979 patients with at least one opioid dispensation in 2017 (Figure 1). This cohort was used to train (n= 275,150, 70%) and validate (n=117,829, 30%) ML models. In 2017 and 2018, 6,608 and 5,423 patients experienced the defined outcome, respectively. Baseline characteristics were different between those who experienced the outcome and those who did not (eTable 3 in Supplement) while characteristics were similar between the training and validation sets (eTable 4 in Supplement). There were 2,283,075 opioid dispensations in 2017 and 1,977,389 in 2018. Overall, in 2017, 2.03% (n= 45,757) of opioid dispensations were associated with the outcome; in 2018, the estimate was 1.6% (n= 31,392).

As described above, we categorized our candidate features into four groups (eTable 5 in Supplement).

**Machine-Learning Prediction Performance**

Using the 2017 validation set, AUROCs for the XGBoost and logistic regression classifiers had the highest discrimination performance at 0.87, while the neural network classifier had lower performance at 0.80 (eTable 6 in Supplement).

Discrimination performance was similar for the 2018 validation set (n=393,023; eTable 6 in Supplement). XGBoost and logistic regression had the highest estimated AUROCs while the neural network classifier was lower (Figure 2A). As expected, precision-recall curves indicate stronger predictive power in opioid dispensations at higher predicted risk percentiles (Figure 2C, 2D).

In the 2018 validation set, although discrimination performance was similar (0.88), individual feature importance was different between the logistic regression and XGBoost classifiers, with logistic regression feature importance more reliant on co-morbidity data from DAD, NACRS and Claims while XGBoost relied more on drug utilization data from PIN (eFigure 2). In the XGBoost classifier, history of drug abuse, alcoholism, and prior hospitalization carried the highest impact for predicting the study outcome (eFigure 3A) where the presence of these features in a patient suggested a strong tendency towards having the defined outcome (eFigure 3B and 3C).

**Calibration**

When considering dispensations predicted to be in the highest percentiles of risk, the top 5-percentile captured 42% of all outcomes using the XGBoost and logistic regression classifiers (Table 1). Also, as the predicted risk percentiles get higher (top 10 percentile to top 0.1 percentile), so too do the corresponding PPVs with the top 0.1 percentile associated with a PPV of 33% for the XGBoost classifier. As well, lower categories of risk percentiles were associated with lower outcomes (Figure 3A and 3B). When we simulated a monitoring workflow scenario with daily data uploads, a similar pattern was illustrated where the dispensations predicted to be higher risk had higher event rates (Figure 3C and 3D).

After using the XGBoost and logistic regression classifiers to identify the dispensations in the highest predicted risk percentiles, the pre-test probability of the outcome (1.6%) was transformed into higher post-test probabilities, with higher probabilities in the riskier percentiles (Table 1). The number needed to screen also decreased as predicted risk increased (Table 1).

Comparing discrimination performance, ML risk prediction outperformed the current guideline approaches when using various combinations of guideline recommendations (Table 2). In many of the guideline scenarios, the estimated AUROCs were close to the 0.5 mark. When we estimated the discrimination performance of the logistic regression classifier based on database source, using all databases produced an AUROC of 0.88. Reducing the database source to only DAD, NACRS, Claims (co-morbidities only) resulted in an AUROC of 0.85, while PIN (prescription history) only was 0.78 (Table 3).

## Discussion

This study showed that ML techniques using available administrative data (prescription histories and ICD codes) may provide enough discriminatory power to predict adverse outcomes associated with opioid prescribing. Indeed, our ML analyses showed very high discrimination performance at 0.88. The linear model (logistic regression) and XGBoosted Trees carried higher discrimination and calibration performance, while the neural network classifier did not perform as well. By identifying the predicted top 5-10 percentile of absolute risk pursuant to an opioid dispensation, we were able to capture approximately half of all outcomes using ML methods. All ML models we trained had higher discrimination performance using independent (external) validation sets than the clinical guideline approach.

Since the prevalence of our defined outcome is relatively low in the general population, PPVs would also be expectedly low. However, estimated PPVs increased when we considered higher risk dispensations, as is expected since PPV is related to event prevalence. This is important because different users of a risk predictor will require different predictive capabilities. Similarly, our estimates of positive likelihood ratios and associated post-test probabilities also increased in dispensations with higher predicted risk indicating the strong predictive power of the XGBoost and logistic regression classifiers; likelihood ratios >10 generate conclusive changes from pre-test to post-test probabilities[32].

The current guideline approach to assess absolute opioid prescribing risk produced c-statistic estimates closer to 0.5 indicating that discrimination was not much better than chance alone. ML models with higher predictive power can better support health departments and PMPs with monitoring mandates to identify and intervene on those at high risk and their associated prescribers. We also found that adding co-morbidity features from administrative

databases increased prediction performance compared to prescription history alone, thus making the case for the use of this data by PMPs and health departments. However, if only prescription history is available, our trained XGBoost classifier still had strong discrimination performance.

We found only one study that used ML approaches to quantify the absolute risk of an event pursuant to an opioid dispensation[10]. Their methodology used rolling 3-month windows for estimating risk and ML model training while we used historic records to estimate 30-day risk. Differences in study population and feature selection may explain why their highest performing ML model was deep learning (neural network classifier) and ours was not. Nevertheless, we were able to replicate and build upon their discrimination performance using our ML approach as we both were able to show that ML approaches have higher predictive power than guideline approaches. Both of our studies used predicted percentile risk estimates to identify high risk dispensations and were able to do so with strong discrimination and calibration performance. This is important because interventions can be targeted to higher risk instead of lower risk patients. Another study we found describes how identifying cases in higher predicted risk percentiles using ML methods can be deployed in hospital settings for the purpose of targeted interventions[30] upon discharge.

The limitations of our study are similar to other ML studies[10] and need to be addressed when considering deployment of ML risk predictors. Our training dataset was not able to account for non-prescription opioid consumption and the risk associated with non-prescription use, both of which are substantial contributors to overall risk[2]. Regarding our analysis, we assumed that all dispensations were independent events; future research in this area should

focus on employing ML methods using correlated data. As with all ML projects, our models were trained using Alberta data and might not be generalizable to other populations, or to specific populations within Alberta. However, our analyses were done on a large population and these results would be expected to be generalizable to the vast majority of patients. Moreover, one of the benefits of the ML process is that models can be retrained or similar methods could be used to develop new models to accommodate different populations.

This study suggests that ML risk prediction can support PMPs, especially if able to use administrative health data. The ML process allows for model training, validation and deployment to specific settings. However, uptake of this technology is limited for the time being. Further research can assess whether implementation of a ML-based monitoring system by PMPs leads to improved clinical outcomes.

## References

1. Belzak L, Halverson J. Evidence synthesis - The opioid crisis in Canada: a national perspective. *Health Promotion and Chronic Disease Prevention in Canada.* 2018;38(6):224-233.
2. Gomes T, Khuu W, Martins D, et al. Contributions of prescribed and non-prescribed opioids to opioid related deaths: population based cohort study in Ontario, Canada. *BMJ.* 2018;362:k3207.
3. Busse JW, Craigie S, Juurlink DN, et al. Guideline for opioid therapy and chronic noncancer pain. *Canadian Medical Association Journal.* 2017;189(18):E659-E666.
4. Dowell D. CDC guideline for prescribing opioids for chronic pain. 2016.
5. ismp Canada. Essential Clinical Skills for Opioid Prescribers. 2017; https://www.ismp-canada.org/download/OpioidStewardship/Opioid-Prescribing-Skills.pdf. Accessed Nov 2018.
6. Centre for Effective Practice. Management of Chronic Non Cancer Pain. 2017; thewellhealth.ca/cncp.
7. College of Physicians and Surgeons of Alberta. TPP ALBERTA MEDICATIONS LIST. *Triplicate Prescription Program* 2020; http://www.cpsa.ca/tpp/tpp-medication-list/. Accessed Jun 2020.
8. World health Organization. Classification of Diseases (ICD). 2019; https://www.who.int/classifications/icd/icdonlineversions/en/. Accessed Jun 2020.
9. Gomes T, Mamdani MM, Dhalla IA, Paterson JM, Juurlink DN. Opioid Dose and Drug-Related Mortality in Patients With Nonmalignant PainOpioid Dose and Drug-related Mortality. *JAMA Internal Medicine.* 2011;171(7):686-691.
10. Lo-Ciganic W-H, Huang JL, Zhang HH, et al. Evaluation of Machine-Learning Algorithms for Predicting Opioid Overdose Risk Among Medicare Beneficiaries With Opioid Prescriptions. *JAMA network open.* 2019;2(3):e190968-e190968.
11. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. *JAMA.* 2017;318(14):1377-1384.
12. Shah NH, Milstein A, Bagley P, Steven C. Making Machine Learning Models Clinically Useful. *JAMA.* 2019;322(14):1351-1352.
13. Liu Y, Chen P-HC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA.* 2019;322(18):1806-1816.
14. Bastanlar Y, Ozuysal M. Introduction to machine learning. *Methods in molecular biology (Clifton, NJ).* 2014;1107:105-128.
15. Thottakkara P, Ozrazgat-Baslanti T, Hupf BB, et al. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PloS one.* 2016;11(5):e0155705.
16. Alberta Machine Intelligence Institute. Machine Learning Process Lifecycle. In:2019.
17. Hsich E, Gorodeski EZ, Blackstone EH, Ishwaran H, Lauer MS. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circulation: Cardiovascular Quality and Outcomes.* 2011;4(1):39-45.
18. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine.* 2015;162(1):W1-W73.
19. equator network. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. 2020; https://www.equator-network.org/reporting-guidelines/tripod-statement/. Accessed Feb 2020.
20. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open.* 2016;6(11):e012799.
21. Government of Canada. Forward Sortation Area—Definition. 2015; https://www.ic.gc.ca/eic/site/bsf-osb.nsf/eng/br03396.html. Accessed April 2020, 2020.

22. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical care.* 2005:1130-1139.
23. World Health Organization. International language for drug utilization research, ATC/DDD. 2020; https://www.whocc.no/. Accessed Jun 2020, 2020.
24. College of Physicians and Surgeons of Alberta. OME and DDD conversion factors. http://www.cpsa.ca/wp-content/uploads/2017/06/OME-and-DDD-Conversion-Factors.pdf.
25. Rose S. Machine Learning for Prediction in Electronic Health Data. *JAMA Network Open.* 2018;1(4):e181404-e181404.
26. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European heart journal.* 2017;38(23):1805-1814.
27. Molnar C. *Interpretable machine learning. A Guide for Making Black Box Models Explainable.* 2019.
28. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Paper presented at: Advances in neural information processing systems2017.
29. Shah ND, Steyerberg EW, Kent DM. Big Data and Predictive Analytics: Recalibrating Expectations. *JAMA.* 2018;320(1):27-28.
30. Morgan DJ, Bame B, Zimand P, et al. Assessment of Machine Learning vs Standard Prediction Rules for Predicting Hospital Readmissions. *JAMA Network Open.* 2019;2(3):e190348-e190348.
31. Centers for Medicare & Medicaid Services (CMS). Announcement of calendar year (CY) 2019 Medicare Advantage capitation rates and Medicare Advantage and Part D payment policies and final call letter.
32. Jaeschke R, Guyatt GH, Sackett DL, et al. Users' Guides to the Medical Literature: III. How to Use an Article About a Diagnostic Test B. What Are the Results and Will They Help Me in Caring for My Patients? *JAMA.* 1994;271(9):703-707.

## Figure Legend

Figure 1. Patient flow diagram of study participants used for training and validating ML models. NACRS: National Ambulatory Care Reporting System; DAD: Discharge Abstract Database; VS: Vital Statistics; PIN: Pharmaceutical Information Network; Claims: Physician Claims

Figure 2. Area under the receiver operating characteristic curve (AUROC) (A) and precision-recall curves (B) for all dispensations using logistic regression, neural network, support vector machine (SVM), XGBoost and Naïve-Bayes; precision-recall curves for higher risk dispensations according to predicted risk percentile categories for logistic regression (C) and XGBoost (D) using the 2018 validation set.

Figure 3. Calibration curves plotting: **1)** observed vs. quantiles of estimated risk for XGBoost (A) and logistic regression (B) classifiers using the 2018 validation dataset and **2)** simulation of a clinical workflow with daily uploads and events per 100 daily dispenses by risk percentiles using 2018 Quarter 1 (Q1) data for logistic regression (C) and XGBoost (D) classifiers. For both classifiers, the majority of counts (dispensations) were predicted to be lower risk.

Table 1. Highest percentiles of estimated risk and predictive power using the XGBoost and logistic regression classifiers for the 2018 validation dataset (n=393,023). Total number of dispenses= 1,977,389; total number of outcomes= 31,392.

| Metric | Top 0.1%ile | | Top 1%ile | | Top 5%ile | | Top 10%ile | |
|---|---|---|---|---|---|---|---|---|
| | XGBoost | Logistic Regression | XGBoost | Logistic Regression | XGBoost | Logistic Regression | XGBoost | Logistic Regression |
| **Number of Dispenses** | 1,977 | 1,977 | 19,774 | 19,774 | 98,869 | 98,869 | 197,739 | 197,739 |
| **TP captured** | 655 | 472 | 4204 | 4100 | 13224 | 13293 | 18404 | 18409 |
| **Percent of TP** | 2.09 | 1.50 | 13.39 | 13.06 | 42.13 | 42.35 | 58.63 | 58.64 |
| **FP captured** | 1322 | 1505 | 15570 | 15674 | 85645 | 85576 | 179335 | 179330 |
| **PPV** | 33.13 | 23.87 | 21.26 | 20.73 | 13.38 | 13.45 | 9.31 | 9.31 |
| **PLR** | 30.71 | 19.44 | 16.74 | 16.22 | 9.57 | 9.63 | 6.36 | 6.36 |
| **Post-test Probability*** | 33.13 | 23.87 | 21.26 | 20.73 | 13.38 | 13.45 | 9.31 | 9.31 |
| **NNS** | 3.17 | 4.49 | 5.08 | 5.22 | 8.48 | 8.43 | 12.95 | 12.95 |

*Pre-test probability estimated at 1.6% using prevalence.

TP: true positives; FP: false positives; PPV: positive predictive value; PLR: positive likelihood ratio; NNS: number needed to screen

Table 2. Discrimination performance of guideline approach using the 2018 validation set. Guideline approaches were adapted from the 2017 Canadian Opioid Prescribing Guideline and 2019 Centers for Medicare & Medicaid Services (CMS) opioid safety measures and compared to logistic regression and XGBoost classifiers (each with an estimated area under the receiver operating characteristic curve of 0.88).

| **Canadian Guidelines** | **AUROC** | **Sensitivity** | **Specificity** |
|---|---|---|---|
| History of mental disorder only | 0.620 | 0.90 | 0.34 |
| Substance abuse only | 0.686 | 0.99 | 0.37 |
| OME/day >90 only | 0.539 | 0.22 | 0.85 |
| (Mental disorder and substance abuse) OR OME/day >90 | 0.690 | 0.91 | 0.47 |
| Mental disorder and substance abuse AND OME/day >90 | 0.560 | 0.20 | 0.91 |
| Mental disorder OR substance abuse OR OME/day >90 | 0.589 | 0.99 | 0.18 |
| **CMS Guidelines** | | | |
| High opioid dose (>120 OME/day for 90+days) | 0.507 | 0.081 | 0.933 |
| Concurrency (Opioid & BZRA for 30+ days) | 0.575 | 0.423 | 0.727 |
| Multiple doctors (>4) | 0.591 | 0.294 | 0.888 |
| Multiple pharmacies (>4) | 0.537 | 0.120 | 0.959 |
| All conditions | 0.50 | 0.001 | 0.999 |
| Any condition | 0.622 | 0.62 | 0.625 |

OME: daily oral morphine equivalents; BZRA: benzodiazepine receptor agonist. Elixhauser scoring ICD codes were used to identify mental disorders and substance abuse.

Table 3. Discrimination performance based on database source using area under the receiver operating characteristic curve (AUROC) for the logistic regression classifier on the 2018 validation set.

| Database source | Predictor Variables formed from database | AUROC |
|---|---|---|
| PIN only | Drug utilization + Prescription history (ATC level 3) | 0.78 |
| DAD, NACRS, Claims | Co-morbidities | 0.85 |
| PIN, DAD NACRS, Claims (all databases used in study) | Demographic + Drug Utilization + Healthcare Utilization + Co-morbidities | 0.88 |

Note: drug utilization includes features describing oral morphine equivalents[24], concurrent use with benzodiazepines, number of opioid and benzodiazepine dispensations, number of unique opioid and benzodiazepine molecules; health care utilization includes features describing number of unique health providers visited, number of hospital visits.

Figure 1. Patient flow diagram of study participants used for training and validating ML models. NACRS: National Ambulatory Care Reporting System; DAD: Discharge Abstract Database; VS: Vital Statistics; PIN: Pharmaceutical Information Network; Claims: Physician Claims

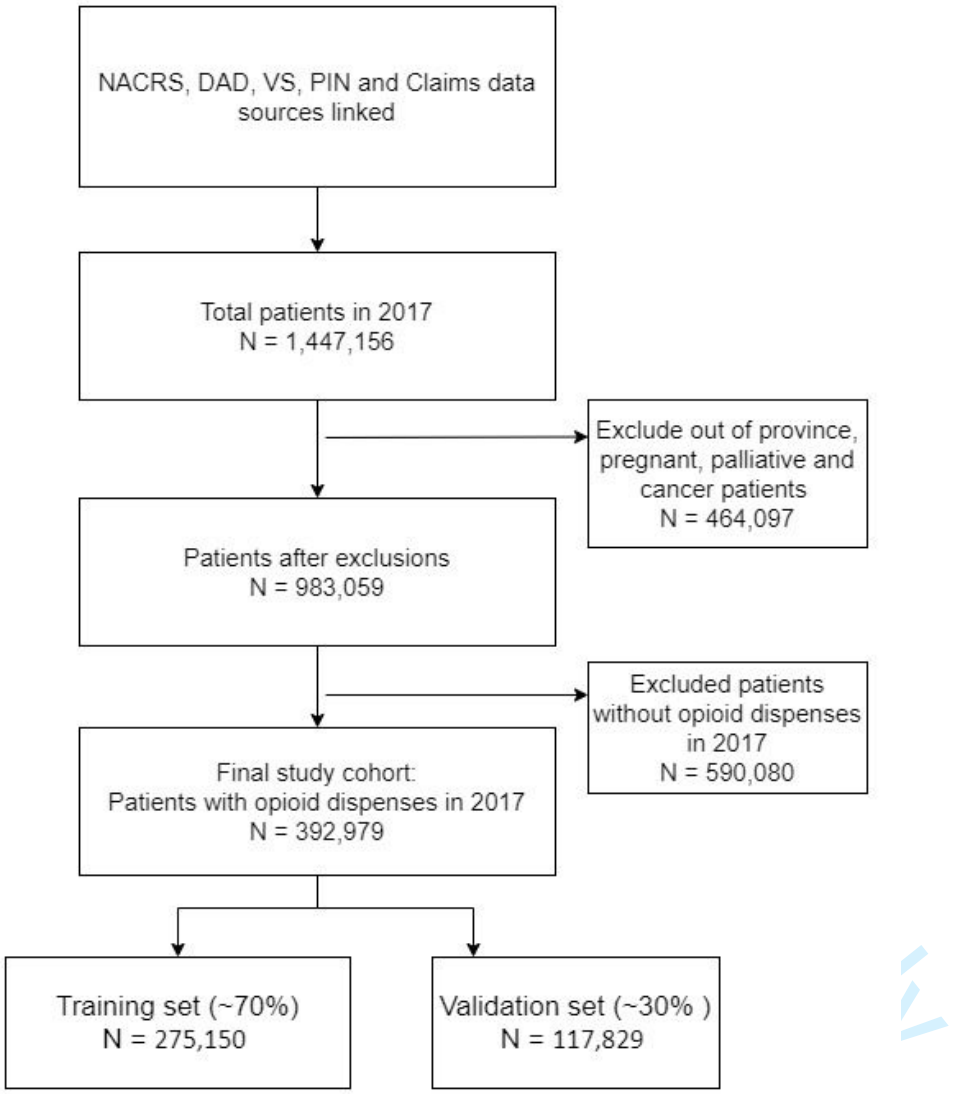
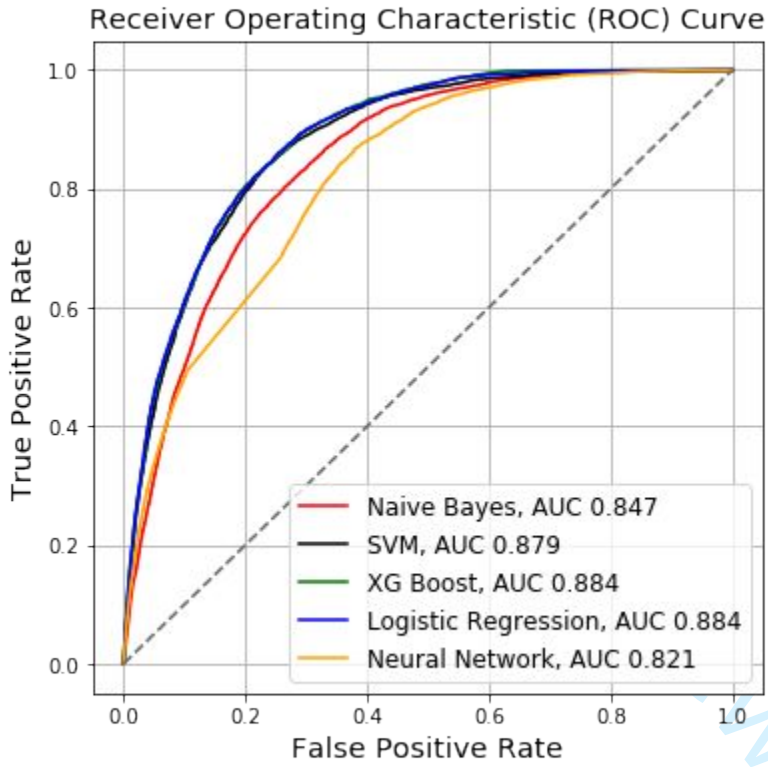
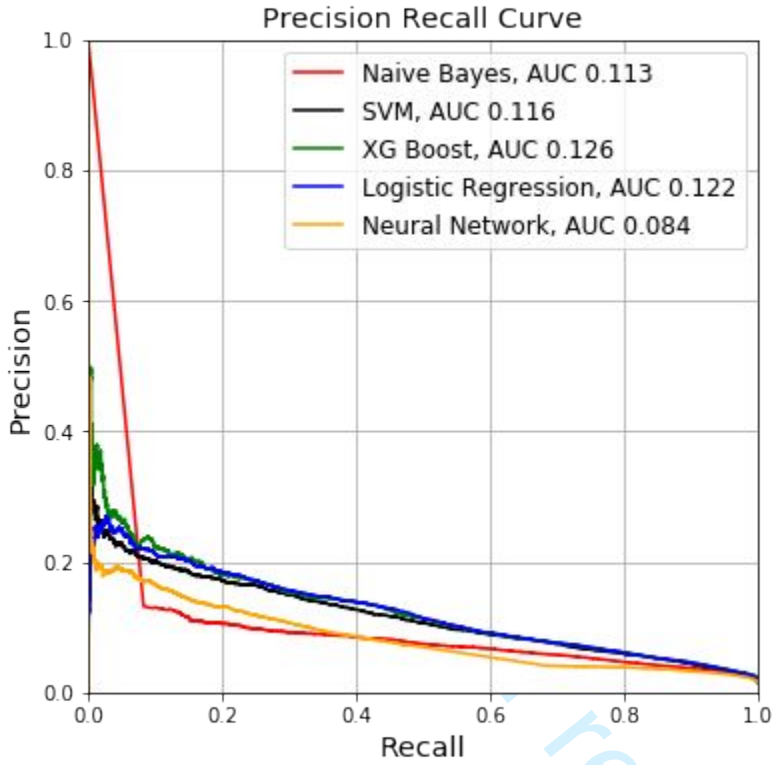


Figure 2. Area under the receiver operating characteristic curve (AUROC) (A) and precision-recall curves (B) for all dispensations using logistic regression, neural network, support vector machine (SVM), XGBoost and Naïve-Bayes; precision-recall curves for higher risk dispensations according to predicted risk percentile categories for logistic regression (C) and XGBoost (D) using the 2018 validation set.

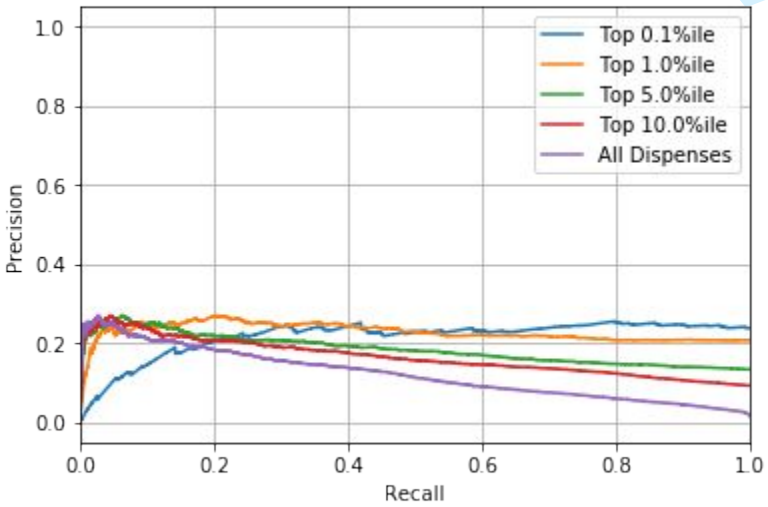
(A)



(B)




(C) Logistic Regression



(D) XGBoost

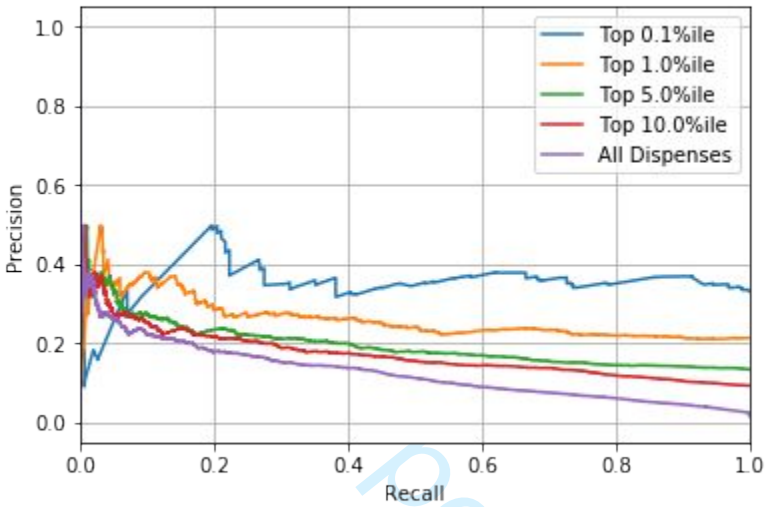
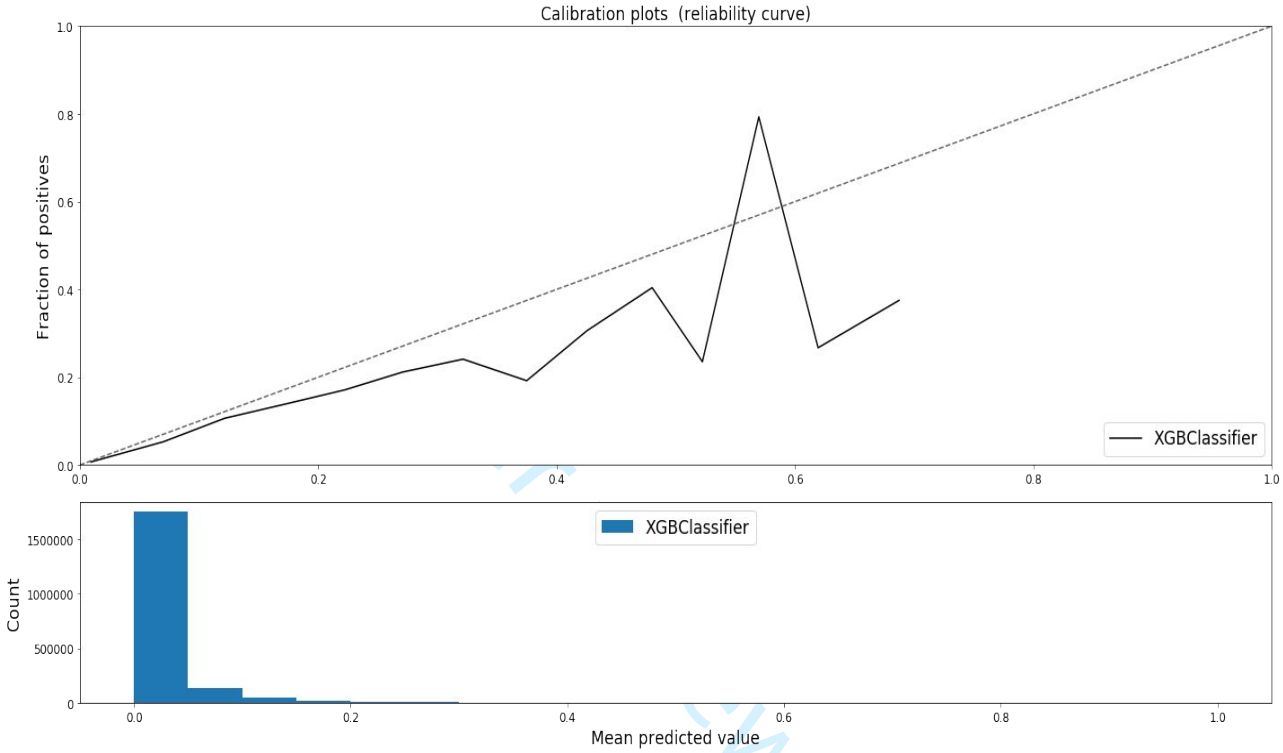
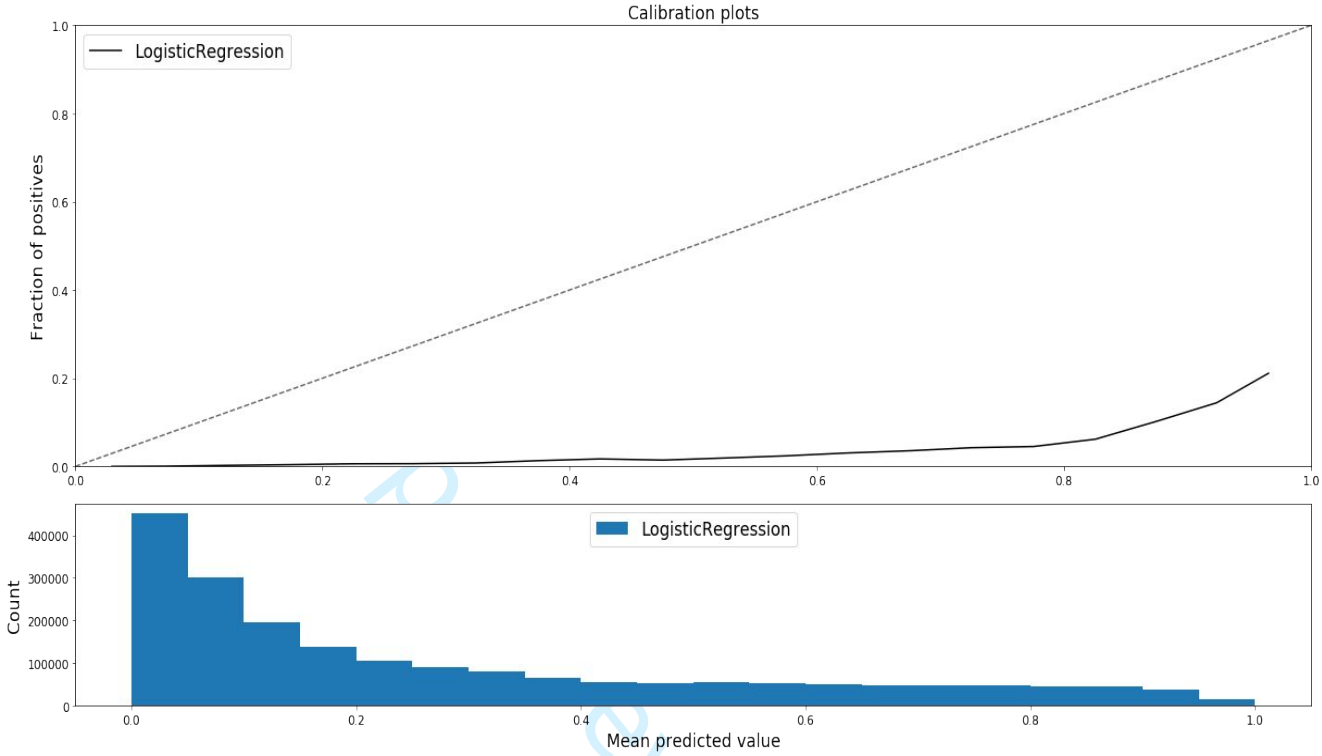


Figure 3. Calibration curves plotting: **1)** observed vs. quantiles of estimated risk for XGBoost (A) and logistic regression (B) classifiers using the 2018 validation dataset and **2)** simulation of a clinical workflow with daily uploads and events per 100 daily dispenses by risk percentiles using 2018 Quarter 1 (Q1) data for logistic regression (C) and XGBoost (D) classifiers. For both classifiers, the majority of counts (dispensations) were predicted to be lower risk.

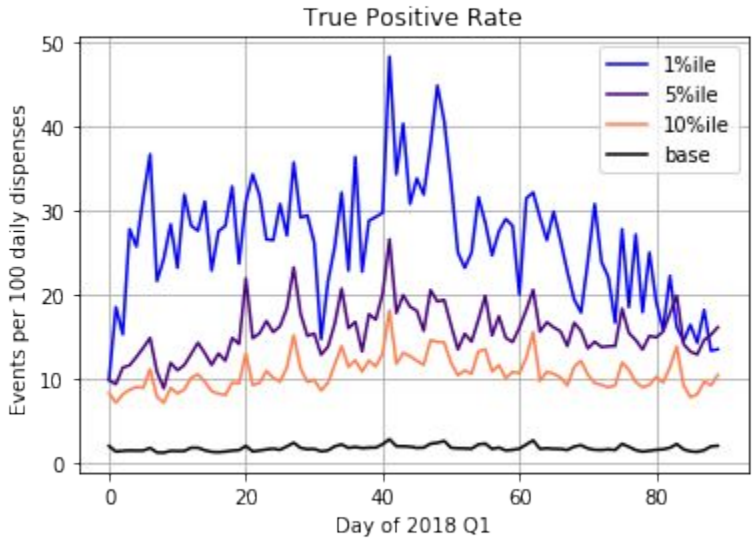
(A) XGBoost



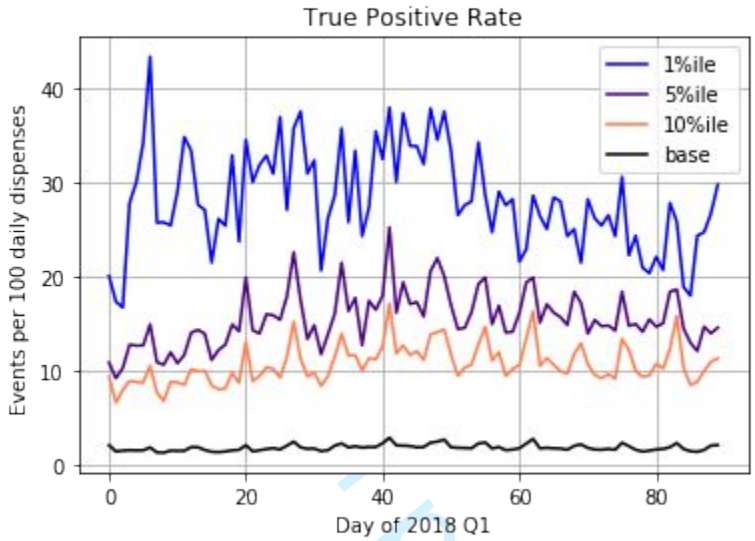
(B) Logistic Regression




(C) Logistic Regression



(D) XGBoost



**Supplementary Content**

**eAppendix.** Machine learning algorithms

**eTable 1.** Diagnostic codes used to exclude patients who had cancer, were pregnant, or were under palliative care.

**eTable 2.** Diagnostic codes used to identify the defined study outcome from emergency visit, hospitalization and death data.

**eTable 3.** Baseline characteristics of study patients (n=392,979). Co-morbidities were determined using Elixhauser criteria. All p-values in the $chi^2$ test of independence were <0.001 unless otherwise indicated.

**eTable 4.** Characteristics of study participants between training and validation groups using 2017 data.

**eTable 5.** Candidate predictors used to train ML algorithms.

**eTable 6.** Discrimination performance using area under the receiver operating characteristic curve (AUROC) of various ML algorithms. Training and validation were done using 2017 data (n=393,979); another independent validation was performed using 2018 data (n=393,023).

**eFigure1.** Schematic of study design and feature generation

**eFigure2**. Feature importance from logistic regression and tree-based (XGBoost) classifiers using the 2018 validation set.

**eFigure3**. Shapley values and feature impact in the XGBoost classifier using the 2018 validation set to describe "associations" between features and the outcome.

**eReferences.**

# eAppendix. Machine Learning Algorithms

## Introduction

While there are always updates and new methods coming up in the fields of machine learning, in this study, we have focused on some of the most reliable and proven approaches for predictive modelling which are explainable and popularly used in previous studies of similar nature.

## Logistic Regression

Regression analysis models the relationship between a dependent variable and a set of independent variables [1]. Typically, this includes understanding how the value of the dependent variable changes with the changes in the values of independent variables. Logistic regression [1] uses the logistic function to model a binary dependent variable, where, based on the values of the independent variables the model can approximate one of the two classes, the instance belongs to. This basic binary model can be extended to deal with multiple classes (e.g. One-vs-all classifiers). However, logistic regression is only capable of modeling a linear relationship of independent variables to the dependent variable, hence limited to problems with linear decision boundaries. We used the sci-kit learn library in our experiments[6] and found L1 regularization to be more effective.

## Ridge Classifier

We used the ridge classifier implemented in the Scikit learn library[5]. It implements a classifier using ridge regression which uses an L2 regularization on the least square objective function. The library converts the labels into -1 and 1 and fits a linear regression on the converted labels with the regularization.

## Random Forest

Random forest is a tree ensemble learning algorithm that has wide applicability in many domains[1]. Random forest is a nonlinear learning algorithm, which arrives at nonlinear decision boundaries by independently combining multiple decision trees. Each individual decision tree in the forest can be grown independently of each other on a subset of the training data. Random forests are mainly sensitive to the number of trees, the depth of a tree and the number of covariates randomly chosen to split at each node[1]. These hyper-parameters can be tuned to find the best configuration of every dataset. Random Forests, in general, are less prone to overfit since they always grow individual trees on a subset of the training data[1]. At prediction time, the decision of each tree is aggregated to compute the final prediction.

## Neural Networks (NN)

Neural networks are another collection of non-linear learning algorithms with high representation power. They are known to be able to find mappings from an input to an output

from a larger non-linear function space [2]. This ability to represent a larger space of nonlinear functions has shown to be very effective recently in many application domains such as natural language processing, computer vision, genomics, computer games and health[2]. Neural networks come in many flavors learning nonlinear mapping of different types of data such as Convolutional NNs being most effective with images and Recurrent NNs for time series and language data. Identifying the most effective neural network structure is one of the difficult and the most time-consuming aspect of applying neural networks to new application domains and data. Generally, neural networks try to exploit the relationships in the raw unstructured data (eg: image and text) presented to the network but with more structured data such as health records and ICD codes learning relationships is much complex. Our neural network models are mainly based on densely connected hidden layers with ReLu[6] activation function. We used the cross-entropy loss for the binary classification Adam optimizer.

**Boosted Learning Algorithms**

Boosting is a process to ensemble multiple base learning algorithms to arrive at better overall performance than any individual base learner[1]. In contrast to independently building multiple models from the subsets of the data, boosting re-weights the training data every time a model is learned for future models. This weighting happens to give more preference to currently misclassified data points in the next round compared to the correctly classified data points. Therefore future learners try to do better on the misclassified data points leading to a collection base learners having a better-combined prediction. This process is sequential so each base learner is dependent on the output of the previously trained model (it is worthy to note XGBoost provides a parallel tree boosting alternative). In our work, we have experimented with several boosting meta-learning algorithms such as XGBoost[7], AdaBoost[5] and GBM[5]. XGBoost uses a variant of trees as the base learner whereas AdaBoost (from Sci-kit learn) can use many ML algorithms as base learners. GBM uses logistic regression by default as the base learner. We used all 3 types of boosting with tuned hyperparameters for comparison.

**Naive Bayes**

Naive Bayes is based on the Bayes theorem with a strong independence assumption between the covariates[1]. This assumption helps in building a simple probabilistic model for learning and inference. Naive Bayes coefficients scale linearly with the number of covariates making this a suitable model for high-dimensional data. We used Naive Bayes as a simple baseline learning algorithm for comparison.

**Support Vector Machines (SVM)**

SVMs[4] are maximum margin classifiers optimizing for learning a hyperplane having the maximum distance away from each of the class data points[1]. SVM is a linear classifier but with the kernel trick to map the inputs to the higher dimensional space, it can learn nonlinear decision boundaries in the input space. SVMs are very effective binary classifiers with the kernel trick[1]. With larger datasets, SVMs tend to become more computationally intensive.

eTable 1. Diagnostic codes used to exclude patients who had cancer, were pregnant, or were under palliative care.

| **Condition** | **ICD 9** | **ICD 10** |
|---|---|---|
| **Cancer** | 140.x - 239.x | C00.x - C99.x, D00.x - D49.x |
| **Pregnancy** | 630.x - 679.x | O00.x - O99.x |
| **Palliative** | V66 | Z51.0, Z51.1, Z51.5 |

eTable 2. Diagnostic codes used to identify the defined study outcome from emergency visit, hospitalization and death data.

| **ICD 10** | **Condition** |
|---|---|
| **T40.x** | Poisoning by, adverse effect of and underdosing of narcotics and psychodysleptics |
| **F55.x** | Abuse of non-psychoactive substances |
| **F11.x - F19.x** | Mental and behavioral disorders due to psychoactive substance use |

eTable 3. Baseline characteristics of study patients (n=392,979). Co-morbidities were determined using Elixhauser criteria. All p-values in the chi² test of independence were <0.001 unless otherwise indicated.

| Characteristic | Number without Event n=386,371 | Percent | Number with Event n=6,608 | Percent |
|---|---|---|---|---|
| **Age:** | | | | |
| Mean (SD) | 48.1 (16.4) | -- | 41.2 (12.4) | -- |
| 18-45 | 162057 | 41.9 | 3466 | 52.4 |
| 45-65 | 154632 | 40.0 | 2656 | 40.2 |
| >65* | 69682 | 18.0 | 486 | 7.4 |
| **Male** | 197491 | 50.3 | 3922 | 59.4 |
| **Female** | 194794 | 49.7 | 2686 | 40.6 |
| **Alcohol Disorder** | 66320 | 16.9 | 5220 | 79.0 |
| **Arrhythmia** | 90621 | 23.1 | 1959 | 29.6 |
| **Blood Loss Anemia** | 1164 | 0.3 | 82 | 1.2 |
| **Congestive Heart Failure** | 18954 | 4.8 | 565 | 8.6 |
| **Coagulopathy** | 8053 | 2.1 | 356 | 5.4 |
| **Deficiency Anemia** | 34188 | 8.7 | 971 | 14.7 |
| **Depression** | 159140 | 40.6 | 5518 | 83.5 |
| **Diabetes** | 64132 | 16.3 | 1408 | 21.3 |
| **Substance Abuse Disorder** | 74678 | 19.0 | 5485 | 83.0 |
| **Fluid Disorder** | 42690 | 10.9 | 3012 | 45.6 |
| **Hypertension** | 140171 | 35.7 | 2624 | 39.7 |
| **Hypothyroidism** | 45519 | 11.6 | 601 | 9.1 |
| **Injury^** | 195688 | 49.9 | 5541 | 83.9 |
| **Liver Disorder** | 21656 | 5.5 | 1588 | 24.0 |
| **Neurologic Disorder** | 230490 | 58.8 | 5387 | 81.5 |
| **Obesity** | 63393 | 16.2 | 970 | 14.7 |
| **Poisoning^** | 17434 | 4.4 | 2775 | 42.0 |
| **Psychoses** | 35870 | 9.1 | 3162 | 47.9 |
| **Renal Disorder** | 16166 | 4.1 | 499 | 7.6 |
| **Rheumatoid Conditions** | 111458 | 28.4 | 3157 | 47.8 |
| **HIV Infection** | 1098 | 0.3 | 141 | 2.1 |
| **Paralysis** | 3874 | 1.0 | 187 | 2.8 |
| **Peptic Ulcer Disease** | 11728 | 3.0 | 509 | 7.7 |
| **Pulmonary Circulation Disorder** | 9611 | 2.4 | 430 | 6.5 |
| **Chronic Pulmonary Disease** | 102990 | 26.3 | 2913 | 44.1 |
| **Peripheral Vascular Disease** | 14467 | 3.7 | 389 | 5.9 |
| **Valvular Disease** | 7308 | 1.9 | 226 | 3.4 |
| **Weight Loss** | 16207 | 4.1 | 747 | 11.3 |

*p-value for age >65 is an estimated 0.037

^ Injury: ICD10: S00-T98; Poisoning: ICD10: T36-T50

eTable 4. Characteristics of study participants between training and validation groups using 2017 data.

| Characteristic | Number in training group N=275,150~ | Percent | Number in validation group N=117,829~ | Percent |
|---|---|---|---|---|
| **Age:** | | | | |
| **Mean (SD)** | 48.3 (16) | -- | 48.2 (16) | -- |
| **18-45** | 114356 | 41.5 | 49909 | 42.3 |
| **45-65** | 111859 | 40.7 | 47132 | 40.0 |
| **>65** | 48935 | 17.8 | 20788 | 17.6 |
| **Male** | 138603 | 48.5 | 59339 | 48.4 |
| **Female** | 136545 | 47.8 | 58490 | 47.7 |
| **Alcohol Disorder** | 46792 | 16.4 | 20199 | 16.5 |
| **Arrhythmia** | 63637 | 22.3 | 27201 | 22.2 |
| **Blood Loss Anemia** | 839 | 0.3 | 336 | 0.3 |
| **Congestive Heart Failure** | 13320 | 4.7 | 5694 | 4.6 |
| **Coagulopathy** | 5697 | 2.0 | 2393 | 2.0 |
| **Deficiency Anemia** | 24096 | 8.4 | 10179 | 8.3 |
| **Depression** | 112080 | 39.2 | 47628 | 38.9 |
| **Diabetes** | 45131 | 15.8 | 19144 | 15.6 |
| **Substance Abuse Disorder** | 52609 | 18.4 | 22713 | 18.5 |
| **Fluid Disorder** | 30272 | 10.6 | 12780 | 10.4 |
| **Hypertension** | 98546 | 34.5 | 41840 | 34.1 |
| **Hypothyroidism** | 31908 | 11.2 | 13666 | 11.2 |
| **Injury*** | 137423 | 48.1 | 58865 | 48.0 |
| **Liver Disorder** | 15252 | 5.3 | 6567 | 5.4 |
| **Neurologic Disorder** | 161706 | 56.5 | 69341 | 56.6 |
| **Obesity** | 44607 | 15.6 | 18882 | 15.4 |
| **Poisoning*** | 12503 | 4.4 | 5293 | 4.3 |
| **Psychoses** | 25422 | 8.9 | 10860 | 8.9 |
| **Renal Disorder** | 11403 | 4.0 | 4817 | 3.9 |
| **Rheumatoid Conditions** | 78268 | 27.4 | 33420 | 27.3 |
| **HIV Infection** | 774 | 0.3 | 336 | 0.3 |
| **Paralysis** | 2717 | 1.0 | 1176 | 1.0 |
| **Peptic Ulcer Disease** | 8239 | 2.9 | 3533 | 2.9 |
| **Pulmonary Circulation Disorder** | 6771 | 2.4 | 2877 | 2.3 |
| **Chronic Pulmonary Disease** | 72265 | 25.3 | 30949 | 25.3 |

| | | | | |
|---|---|---|---|---|
| **Peripheral Vascular Disease** | 10228 | 3.6 | 4278 | 3.5 |
| **Valvular Disease** | 5111 | 1.8 | 2215 | 1.8 |
| **Weight Loss** | 11477 | 4.0 | 4790 | 3.9 |

~p-values for chi$^2$ test of independence were all >0.06 when comparing training and validation sets.

*Injury: ICD10: S00-T98; Poisoning: ICD10: T36-T50

eTable 5. Candidate predictors used to train ML algorithms.

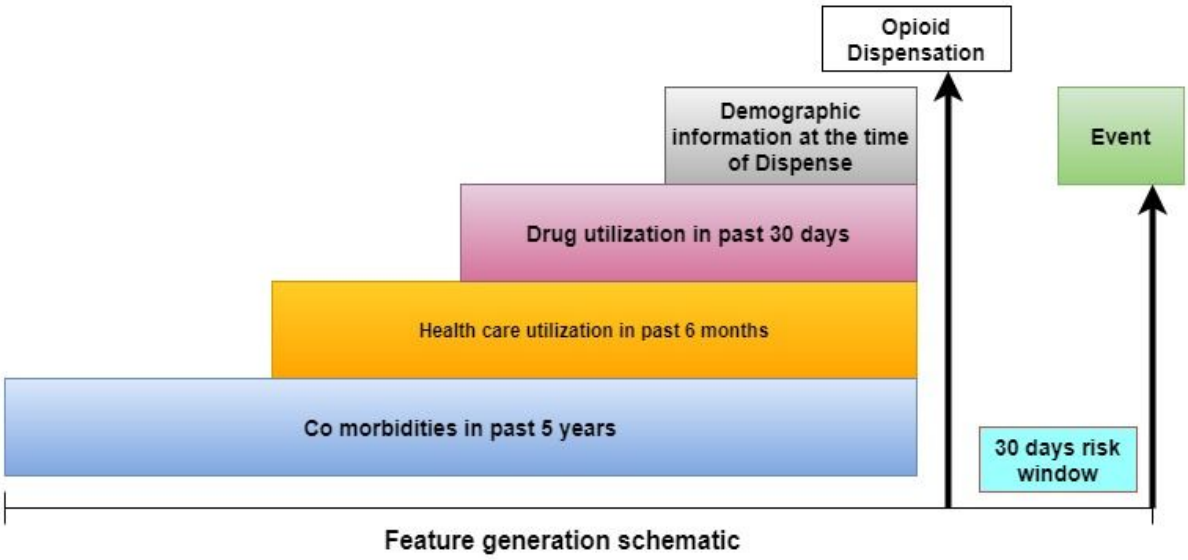
| **Category (data source)** | **Description** |
|---|---|
| Demographic information (PIN) | age, sex, postal codes, mean income |
| Drug utilization history (PIN) | drug dispenses in past 30 days using on ATC codes, oral morphine equivalents, concurrent use with benzodiazepines, number of dispensations and unique molecules of opioids and benzodiazepines |
| Health care utilization (PIN DAD) | flags for previous hospitalizations, number of unique providers |
| ICD based co-morbidities (DAD, NACRS, Claims) | Elixhauser condition flags based on the past 5 years of claims, hospitalizations, and emergency visits. |

Note: ICD: International Statistical Classification of Diseases and Related Health Problems, World Health Organization.

eTable 6. Discrimination performance using area under the receiver operating characteristic curve (AUROC) of various ML algorithms using all features (demographics, health utilization, prescription history, co-morbidities). Training and validation were done using 2017 data (n=393,979); another independent validation was performed using 2018 data (n=393,023).

| Algorithm | Train | Validation 2017 | Validation 2018 |
|---|---|---|---|
| XGBoost Classifier | 0.897 | 0.870 | 0.884 |
| Logistic Regression | 0.887 | 0.869 | 0.884 |
| Gradient Boosting Classifier | 0.898 | 0.868 | 0.883 |
| AdaBoost Classifier | 0.884 | 0.868 | 0.882 |
| Random Forest Classifier | 0.909 | 0.863 | 0.881 |
| Ridge Classifier | 0.895 | 0.863 | 0.879 |
| SVM | 0.896 | 0.860 | 0.878 |
| Gaussian Naive Bayes | 0.846 | 0.826 | 0.847 |
| Decision Tree Classifier | 0.919 | 0.791 | 0.822 |
| Neural Networks | 0.827 | 0.804 | 0.821 |

**eFigure 1. Schematic of study design and feature generation**



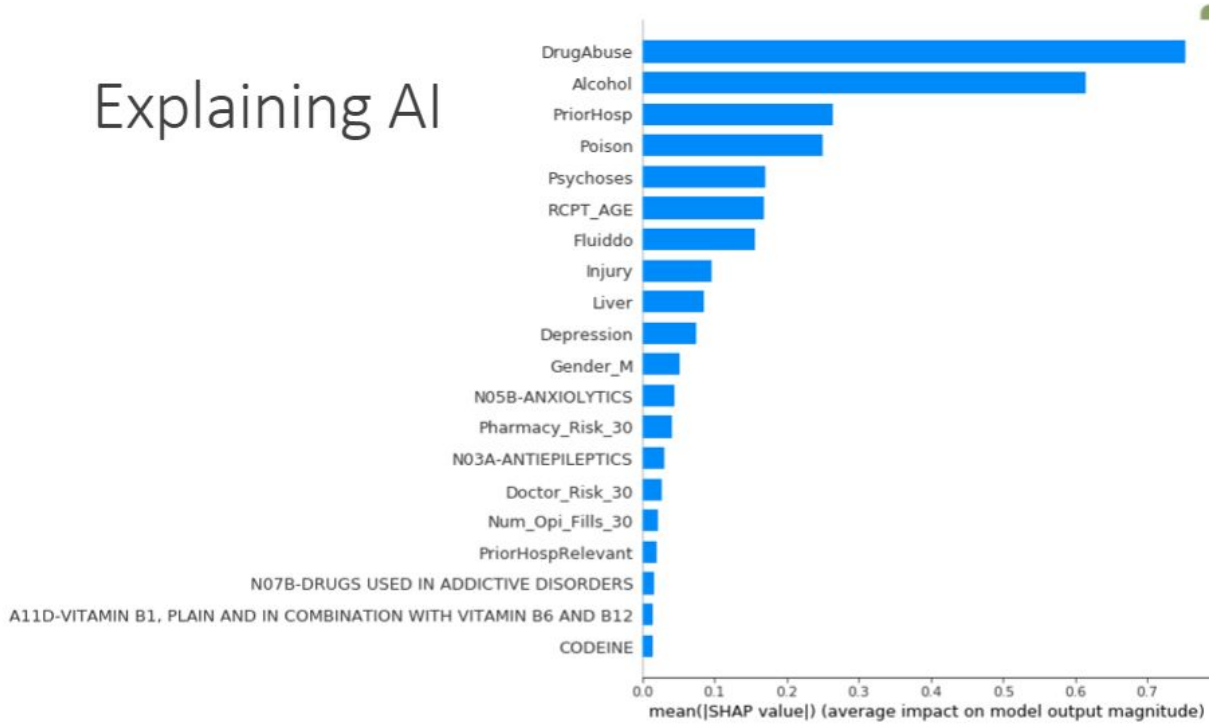
**eFigure2. Feature importance from logistic regression and tree-based (XGBoost) classifiers using the 2018 validation set. The logistic regression classifier relied more on co-morbidity data from DAD, NACRS, and Claims databases; XGBoost classifier relied more on data from the PIN database. AUROCs for both classifiers were similar at 0.88.**

| LOGISTIC REGRESSION | |
|---|---|
| Drug Abuse | 1.00 |
| Age | 0.65 |
| Prior Hospitalization | 0.62 |
| Alcohol Abuse | 0.62 |
| Fluid Disorder | 0.32 |
| Substance Poison | 0.31 |
| Psychoses | 0.31 |
| Num_Benzo_Ingred_30 | 0.26 |
| Depression | 0.19 |
| Concurrent_Opioid_Benzo_30 | 0.19 |
| Injury | 0.17 |

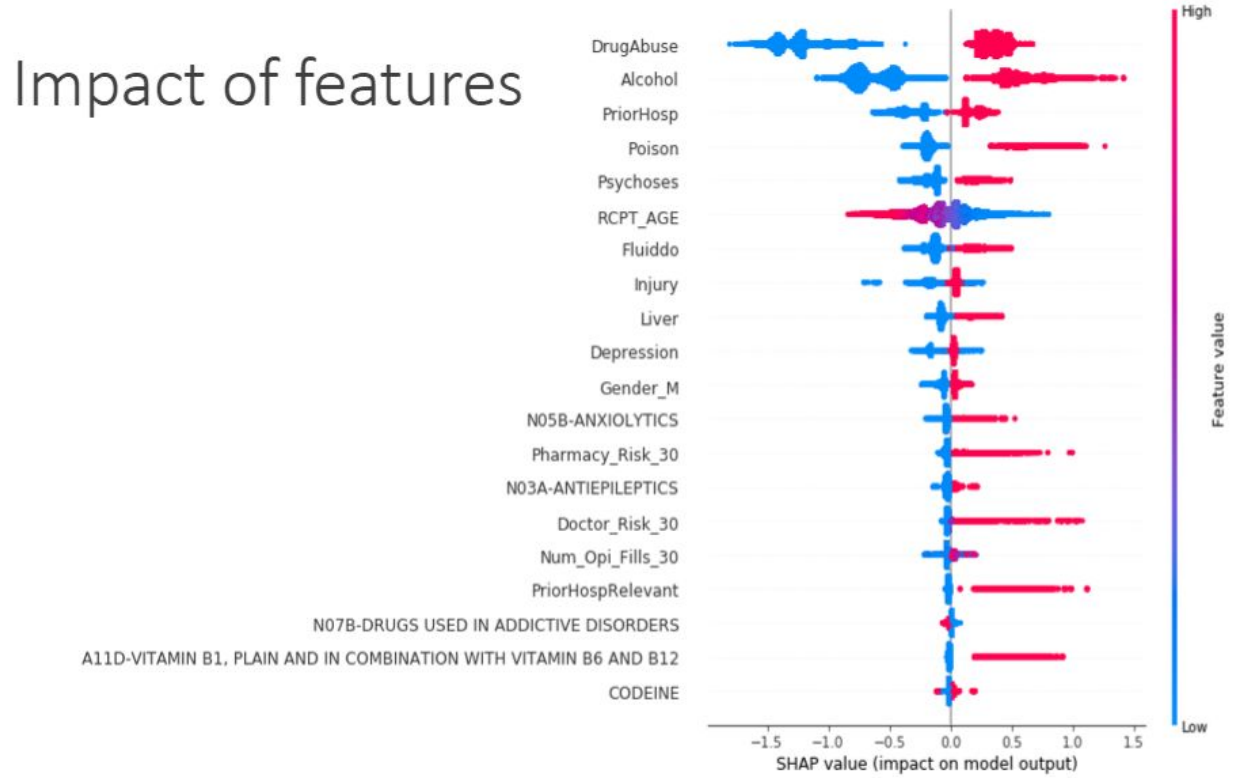
| TREE-BASED MODEL | |
|---|---|
| Age | 1.00 |
| Num_Fills_30 | 1.00 |
| Num_Opioid_Fills_30 | 0.86 |
| Num_Benzo_Fills_30 | 0.46 |
| Doctor_Risk_30 | 0.45 |
| Total_OME_30_Days_Supply | 0.43 |
| Substance poison | 0.37 |
| Pharmacy_Risk_30 | 0.35 |
| Num_Doctors_30 | 0.34 |
| Income | 0.34 |
| Prior hospitalization | 0.26 |

**eFigure 3. Shapley values and feature impact in the XGBoost classifier using the 2018 validation set to describe "associations" between features and the outcome. Features with the most impact on the model with drug abuse with drug abuse ranked highest (A); tornado plot illustrating feature impact. Red indicates higher impact and plots to the right of 0.0 indicate the tendency to be associated with the study outcome while blue indicates lower impact and plots to the left of 0.0 indicate the tendency to be associated with no outcome (B); explaining the prediction of study outcomes based on predictor values for 4 patients (C).**

**(A)**

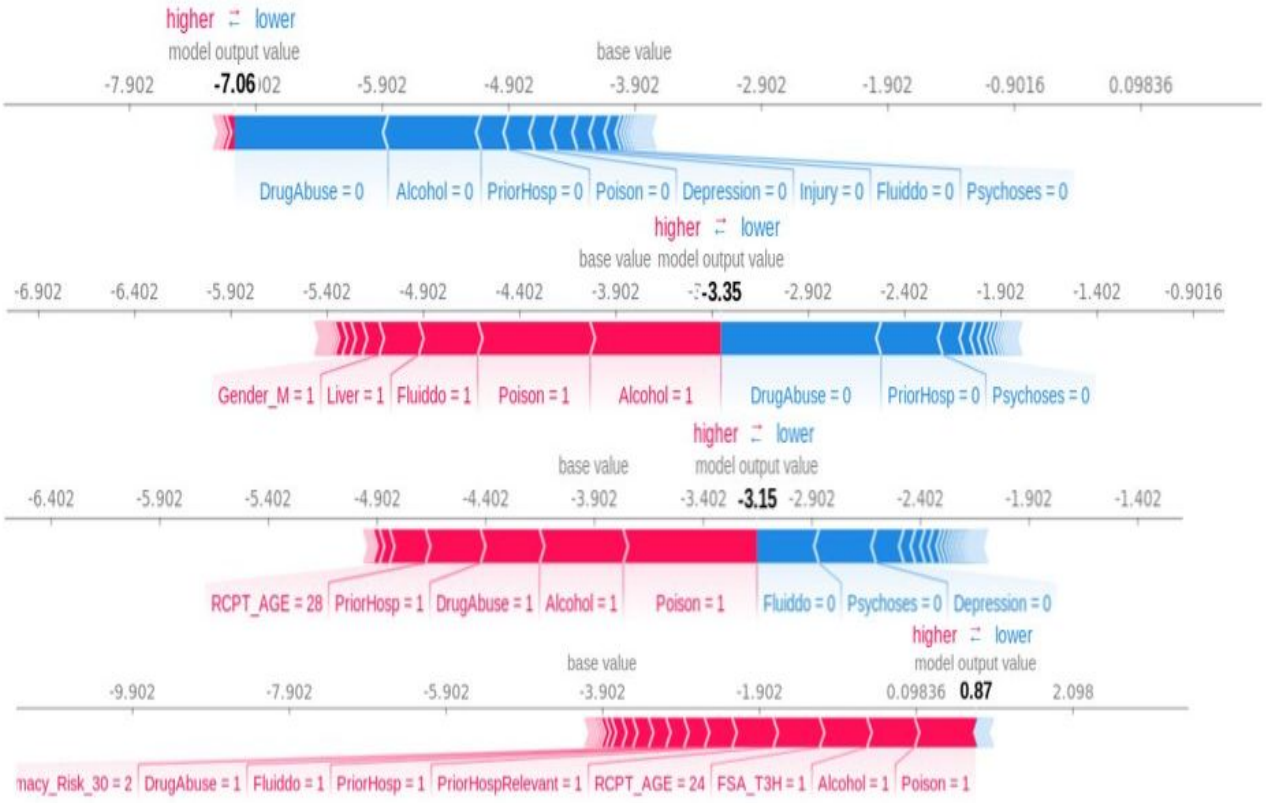


**(B)**




Note: RCPT_AGE- age at opioid dispensation; Fluiddo- fluid disorder according to Elixhauser co-morbidity; Gender_M-male sex' NO5B-ANXIOLYTICS- prescribed ATC code benzodiazepine derivatives; Pharmacy_Risk_30- derived feature using proportion of opioid/benzodiazepine patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each pharmacy; N03A-ANTIEPILEPTICS- ATC code for anti-epileptics dispensed to patient; Doctor_Risk_30- derived feature using proportion of opioid/benzodiazepine patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each physician; Num_Opi_Fills_30- number of opioid dispensations in the previous 30 days prior to opioid dispensation; PriorHospRelevant- flag for history of any opioid related hospitalization in the previous 180 days prior to opioid dispensation; N07B-DRUGS USED IN ADDICTIVE DISORDERS- ATC code for drugs dispensed to patient for treating substance abuse disorders; A11D-VITAMIN B1, PLAIN AND IN COMBINATION WITH VITAMIN B6 AND B12- ATC code for patients dispensed Vitamins B1, B6, or B12; CODEIN: history of codeine use

**(C)**




**Note:** The "reference point" is called the "base value" at -3.902. Values in bold to the left of the base value indicate a lower predicted probability of the study outcome and values in bold to the right indicate a higher predicted probability of the study outcome. The top plot describes a patient at "low risk" for the study outcome. As can be seen from the feature values, this patient has a negative history for the specified features. The middle 2 plots describe a patient at "medium risk" while the bottom plot shows a patient at "high risk" for the study outcome.

**eReferences**


1. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning, vol. 1. Springer series in statistics New York (2001)
2. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
3. Zhu, H. Zou, S. Rosset, T. Hastie, “Multi-class AdaBoost”, 2009.
4. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2011 May 6;2(3):1-27.
5. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
6. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. InProceedings of the 27th international conference on machine learning (ICML-10) 2010 (pp. 807-814).
7. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).

BMJ Open

# Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada

| | |
|---|---|
| Journal: | *BMJ Open* |
| Manuscript ID | bmjopen-2020-043964.R1 |
| Article Type: | Original research |
| Date Submitted by the Author: | 26-Jan-2021 |
| Complete List of Authors: | Sharma, Vishal; University of Alberta, School of Public Health<br>Kulkarni, Vinaykumar; OKAKI Health Analytics<br>Eurich, Dean; University of Alberta, School of Public Health<br>Kumar, Luke; Alberta Machine Intelligence Institute<br>Samanani, Salim; Okaki Health Intelligence, |
| Primary Subject Heading: | Epidemiology |
| Secondary Subject Heading: | Health informatics, Public health |
| Keywords: | PUBLIC HEALTH, EPIDEMIOLOGY, Adverse events < THERAPEUTICS, Health & safety < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Clinical governance < HEALTH SERVICES ADMINISTRATION & MANAGEMENT |
| | |

*I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.*

*The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.*

*Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.*

**Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada**

**Author list (in order):**

Vishal Sharma (0000-0001-7907-1183), Vinaykumar Kulkarni, Dean T. Eurich (0000-0003-2197-0463), Luke Kumar, Salim Samanani (0000-0001-6751-4805)

**Address for each author:**

2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of Alberta, Edmonton, Alberta, Canada, T6G 2E1 Vishal Sharma BPharm PhD Candidate,

OKAKI Health Intelligence, Edmonton, Alberta, Canada, Vinaykumar Kulkarni MSc

2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of Alberta, Edmonton, Alberta, Canada, T6G 2E1 Dean Eurich professor

Alberta Machine Intelligence Institute, Edmonton, Alberta, Canada, T5J 3B1 Luke Kumar MSc

OKAKI Health Intelligence, Calgary, Alberta, Canada, Salim Samanani MD, Medical Director

**Corresponding Author:**

Dean Eurich, 2-040 Li Ka Shing Center for Health Research Innovation, University of Alberta, Edmonton, Alberta, Canada, T6G 2E1; Phone 780-492-6333; fax 780-492-7455; email: deurich@ualberta.ca

**Acknowledgement**

This study is based on data provided by The Alberta Strategy for Patient Orientated Research (AbSPORU) SUPPORT unit and Alberta Health. The interpretation and conclusions contained herein are those of the researchers and do not necessarily represent the views of the Government of Alberta or AbSPOR. Neither the Government of Alberta, ABSPOR nor Alberta Health expresses any opinion in relation to this study. This work was supported by Mitacs through the Mitacs Accelerate Program (VS and DTE).

**Contributors:** VS VK LK SS and DTE were involved in the conception and design of the study. VS VK LK SS and DTE analyzed the data. VS VK and LK drafted the article. VS VK LK DTE and SS revised the article. All authors gave final approval of the version to be published. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. DTE is the guarantor.

**Funding:** This study received no funding.



***Competing Interest:*** *All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; Salim Samanani has received grants from the College of Physicians & Surgeons of Alberta; no other relationships or activities that could appear to have influenced the submitted work.*

**Ethical approval:** This study was approved by the Health Research Ethics board at the University of Alberta (#Pro00083807_AME2).

**Data Sharing:** The data used in this study is not available for external analysis. However, administrative health data can be accessed from Alberta Health by following defined research protocols and confidentiality agreements.

**Transparency:** The lead author, VS, (the manuscript's guarantor, Dean Eurich) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as originally planned (and, if relevant, registered) have been explained.

**Word Count: 3120**

## Abstract

**Objective:** To develop machine-learning models employing administrative-health data that can estimate risk of adverse outcomes within 30-days of an opioid dispensation for use by health-departments or prescription monitoring programs.

**Design, Setting, and Participants:** This prognostic study was conducted in Alberta, Canada between 2017-2018. Participants included all patients over 18 years of age who received at least one opioid dispensation. Pregnant and cancer patients were excluded.

**Exposure:** Each opioid dispensation served as an exposure.

**Main Outcomes/Measures:** Opioid related adverse outcomes were identified from linked administrative health-data. Machine-learning algorithms were trained using 2017 data to predict risk of hospitalization, emergency department visit, and mortality within 30-days of an opioid dispensation. Two validation sets, using 2017 and 2018 data, were used to evaluate model performance. Model discrimination and calibration performance were assessed for all patients and those at higher risk. Machine-learning discrimination was compared to current opioid guidelines.

**Results:** Participants in the 2017 training set (n=275,150) and validation set (n=117,829) had similar baseline characteristics. In the 2017 validation set, c-statistics for the XGBoost, logistic regression, and neural network classifiers were 0.87, 0.87, and 0.80, respectively. In the 2018 validation set (n=393,023), the corresponding c-statistics were 0.88, 0.88, and 0.82. C-statistics from the Canadian guidelines ranged from 0.54-0.69 while the US guidelines ranged from 0.50-0.62. The top 5-percentile of predicted risk for the XGBoost and logistic regression classifiers captured 42% of all events and translated into post-test probabilities of 13.38 and 13.45%, respectively, up from the pre-test probability of 1.6%.

**Conclusion:** Machine-learning classifiers, especially incorporating hospitalization/physician claims data, have better predictive performance compared to guideline or prescription history only approaches when predicting 30-day risk of adverse outcomes. Prescription monitoring programs and health departments with access to administrative data can use machine-learning classifiers to effectively identify those at higher risk compared to current guideline-based approaches.

## Article Summary

### Strengths and Limitations:

- This study incorporated near complete capture of opioid dispensations from community pharmacies and used validated administrative health data.
- The study population is the entire provincial population and is generalizable to other populations in Canada and beyond.
- This study used commonly available algorithms to train machine-learning models using data which is available to government health departments in all provinces in Canada and other single payer jurisdictions; ML classifiers were evaluated with informative prognostic metrics not usually seen in other studies like ours.
- Our predictive models used dispense events and not medication utilization, which is difficult to capture in administrative data.
- Our training dataset does not account for non-prescription opioids, opioids administered in hospitals, and other risks associated with non-prescription use.

## Introduction

Canada is among the countries with the highest rates of opioid prescribing in the world, making prescription opioid use a key driver of the current opioid crisis[1]; a major part of the policy response to the opioid crisis focuses on endorsing safe, appropriate opioid prescribing[2-4]. In order to minimize high risk opioid prescribing and to identify patients at high risk of opioid related adverse outcomes, numerous health regulatory bodies have released clinical practice recommendations for health providers regarding appropriate opioid prescribing[3,5,6].

Prescription monitoring programs (PMPs) have been implemented around the world, like Alberta's provincial Triplicate Prescription Program (TPP)[7] in Canada, and are mandated to monitor the utilization and appropriate use of opioids to reduce adverse outcomes. In most jurisdictions, both population-level monitoring metrics and clinical decision aids are used to identify patients at risk of hospitalization or death and are most often based on prescribing guidelines. However, a comprehensive infrastructure of administrative data containing patient level ICD[8] codes and prescription drug histories exists in Alberta and other provinces in Canada which could be further integrated to predict opioid-related risk. Furthermore, current guidelines' of high risk prescribing and utilization of opioids were derived from studies that used traditional statistical methods (regression analyses) to identify population level risk factors for overdose rather than an individual's absolute risk[3,9,10]; these population estimates may not be generalizable to different populations[11]. Thus, a functional gap exists in many health jurisdictions where much of the available administrative health data is not being leveraged for opioid prescription monitoring.

Supervised machine learning (ML)[12,13] is an approach that uses computer algorithms to build predictive models in the clinical setting that can make use of the large amounts of available administrative data[14,15], all within a well-defined process[16]. Supervised ML trains on labelled data to develop prediction models that are specific to different populations and, in many cases, can provide better predictive performance than traditional, population-based statistical models[10,15,17]. We identified one study[10] that applied ML techniques to predict overdose risk in opioid patients pursuant to a prescription. In their validation sample, they found that the DNN (deep neural network) and GBM (gradient boosting machine) algorithms carried the best discrimination performance based on estimated c-statistics and that the ML approach out-performed the guideline approach in terms of risk prediction; neural networks have little interpretability and are not necessarily better at predicting outcomes when trained on structured data[18]. This study relied on c-statistics to evaluate their ML models and did not emphasize other performance metrics required to assess clinical utility that are recommended by medical reporting guidelines[11,13,19,20]. It also did not address the important issue of ML model interpretability[21]. Reporting informative prognostic metrics is needed to better understand the capabilities of ML classifiers if health departments and PMPs are to incorporate them into their decision-making processes.

The objective of our study was to further develop and validate ML algorithms (beyond just DNN) to predict the 30-day risk of hospitalization, emergency visit and mortality for a patient in Alberta, Canada at the time of an opioid dispensation using administrative data routinely available to health departments and PMPs and evaluate them using the above referenced reporting guidelines. We also analyzed feature importance to provide meaningful

interpretations of the ML models. Comparing discrimination performance (area under the receiver operating characteristics curves), we hypothesized that the ML process would perform better than the current guideline approach for predicting risk of adverse outcomes related to opioid prescribing.

## Methods

### Study Design and Participants

This prognostic study used a supervised ML scheme. All patients in Alberta, Canada who received a dispensation for an opioid, were 18 years of age and older between Jan 1, 2017 and Dec 31, 2018 were eligible. Patients were excluded from all analyses if they had any previous diagnosis of cancer, received palliative interventions or were pregnant during the study period (eTable 1 in Supplement) as use of opioids in these contexts is clinically different.

Government health departments and payers in many jurisdictions have systems to capture prescription histories and ICD diagnostic codes. As such, we linked various administrative health data sets available in Alberta, Canada using unique patient identifiers in order to establish a complete description of patient demographics, drug exposures and health outcomes. These databases include *1) Pharmaceutical Information Network (PIN):* PIN data includes all dispensing records from community pharmacies from all prescriber types occurring in the province outside of the hospital setting. PIN collects all drug dispensations irrespective of age or insurance status in Alberta, 2*) Population and Vital Statistics Data (VS, Alberta Services)*: sex, age, date of birth, death date, immigration and emigration data, and underlying cause of death according to the World Health Organization algorithm using ICD codes[8], *3)*

*Hospitalizations and Emergency Department Visits (NACRS [National Ambulatory Care Reporting System], DAD [Discharge Abstract Database]):* all services, length of stay, diagnosis (up to 25 ICD-10[8] based diagnoses). Data and coding accuracy are routinely validated both provincially and centrally via the Canadian Institute for Health Information, and 4*) Physician Visits/Claims (Alberta Health):* all claims from all settings (e.g., outpatient, office visits, emergency departments, inpatient) with associated date of service, ICD code, procedure and billing information.

This study followed the TRIPOD and STARD reporting guidelines[22-24] and received ethics approval from the University of Alberta ethics board (Pro00083807_AME2). All analyses were done using Python (v. 3.6.8,), SciKit Learn[25] (v. 0.23.2) SHAP[26] (v. 0.35), XGBoost (v. 0.90)[27], Pandas (v. 1.0.5)[28] and H20 Driverless AI (version 1.9).

**Measures and Outcome**

ML models were trained on a labelled dataset in which the observation/analysis unit was an opioid dispensation. The primary outcome was a composite of a drug-related hospitalization, emergency department (ED) visit or mortality within 30 days of an opioid dispensation based on ICD-10 codes (T40, F55, F10-19; eTable 2 in Supplement)[2,10,29].

We anticipated that our defined outcome would be a rare event, leading to a class imbalanced dataset[30]. To address this, we relied on specifying balanced class weightage for supporting algorithms; other approaches were not deemed suitable (e.g., randomly repeating minority class) and under sampling (sub-sampling within the majority class) resulted in changes in outcome prevalence.

**Predictor Candidates for ML Models**

Predictor variables in our ML models included those that were informed by the literature[3,4,10] and those directly obtained from the data sets. These included features based on demographics (age, sex, income using Forward Sortation index from postal codes[31]), co-morbidity history using ICD-based Elixhauser score categories[32], health care utilization (number of unique opioid prescribers, number of hospital visits), and drug utilization (level 3 ATC codes[33], oral morphine equivalents[34], concurrent use with benzodiazepines, number of opioid and benzodiazepine dispensations, number of unique opioid and benzodiazepine molecules). Depending on the potential predictor and data availability, we used data from 30 days to 5 years before the opioid dispensation to generate model features (eFigure 1 in Supplement); 30 days was used to reflect the immediate nature of the risk and 5 years to fully capture co-morbidities. This approach aligns with how health providers would assess patients using the entire history of co-morbidities and then the more immediate factors in deciding on the need for a therapeutic as well as risk in patients. Experiments were performed to identify the features and data sets that contributed most to predicting the outcomes, with a view to minimizing the potential future data requirements for health departments and PMPs.

**Statistical Analyses and Machine-Learning Prediction Evaluation**

We randomly divided the patients in the 2017 portion of our study cohort into training (70%) and validation (30%) sets[13] by patients and opioid dispensations such that no patients in the training set were in the validation set. Baseline characteristics and event rates were compared in the training vs validation group, and between those who experienced the outcome and those

who did not, using chi-squared tests of independence. As well, we used all 2018 data as another independent validation set.

First, we trained commonly used[13,35] ML algorithms (eAppendix in Supplement) and out of box models were further tuned when training on the dataset using 5-fold cross validation on the training data. to address model overfitting[13,36]. As is common in ML validation studies[10,13], we reported model discrimination performance (i.e. how well a model differentiates those at higher risk from those at lower risk)[11] using area under the receiver operating characteristic curve (AUROC; c-statistic). We then stratified the two ML models with the highest c-statistics into percentile categories according to absolute risk of our outcome, as was done in previous studies[10,37]. We also plotted AUROC[11] and precision-recall curves (PRCs)[38].

Because discrimination alone is insufficient to assess ML model prediction capability, we assessed a second necessary property, namely, calibration (i.e., how similar the predicted absolute risk is to the observed risk across different risk strata)[11,39]. Using the two ML models with the highest discrimination performance discussed above, we assessed calibration performance on the 2018 data by plotting observed (fraction of positives) vs predicted risk (mean predicted value). Using these two ML classifiers, we analyzed the top 0.1, 1, 5, and 10 percentiles of predicted risk by the number of true and false positives, positive likelihood ratios (PLR)[20], post-test probabilities, and number needed to screen. We also performed a simulation of daily data uploads for 2018 Quarter 1 to view the predictive capabilities if a ML risk predictor were to be deployed into a monitoring workflow.

For the XGBoost and logistic regression classifiers, we reported feature importance[35] and plotted PRCs that compared all dispenses to those within the top 10 percentiles of estimated risk. As well, for the XGBoost classifier, we described feature importance on model outcome using SHAP values[26,40] to add an additional layer of interpretability.

Finally, we compared ML risk prediction (the two ML models with highest discrimination performance) to current guideline approaches as others have[10], using the 2019 Centers for Medicare & Medicaid Services (CMS) opioid safety measures[41] and the 2017 Canadian Opioid Prescribing Guideline[3]. We also compared the discrimination performance of different logistic regression classifier models using various combinations of features derived from their respective databases: **1)** demographic and drug/health utilization features from PIN and 2**)** co-morbidity features derived from DAD, NACRS and Claims.

### Patient and Public Involvement

This research was done without patient involvement. Patients were not invited to comment on the study design and were not consulted to develop patient relevant outcomes or interpret the results. Patients were not invited to contribute to the writing or editing of this document for readability or accuracy. There are no plans to disseminate the results of the research to study participants.

## Results

### Patient Characteristics and Predictors

We identified 392,979 patients with at least one opioid dispensation in 2017 (Figure 1). This cohort was used to train (n= 275,150, 70%) and validate (n=117,829, 30%) ML models. In 2017 and 2018, 6,608 and 5,423 patients experienced the defined outcome, respectively. Baseline characteristics were different between those who experienced the outcome and those who did not (eTable 3 in Supplement) while characteristics were similar between the training and validation sets (eTable 4 in Supplement). There were 2,283,075 opioid dispensations in 2017 and 1,977,389 in 2018. Overall, in 2017, 2.03% (n= 45,757) of opioid dispensations were associated with the outcome; in 2018, the estimate was 1.6% (n= 31,392).

As described above, we categorized our candidate features into four groups (eTable 5 in Supplement).

**Machine-Learning Prediction Performance**

Using the 2017 validation set, AUROCs for the XGBoost and logistic regression classifiers had the highest discrimination performance at 0.87, while the neural network classifier had lower performance at 0.80 (eTable 6 in Supplement).

Discrimination performance was similar for the 2018 validation set (n=393,023; eTable 6 in Supplement). XGBoost and logistic regression had the highest estimated AUROCs and area under PRCs while the neural network classifier was lower (Figure 2A, 2B). As expected, precision-recall curves indicate stronger predictive performance in opioid dispensations at higher predicted risk percentiles (Figure 2C, 2D).

In the 2018 validation set, although discrimination performance was similar (0.88), individual feature importance was different between the logistic regression and XGBoost

classifiers, with logistic regression feature importance more reliant on co-morbidity data from DAD, NACRS and Claims while XGBoost relied more on drug utilization data from PIN (eFigure 2). In the XGBoost classifier, history of drug abuse, alcoholism, and prior hospitalization carried the highest importance for predicting the study outcome (eFigure 3A) where the presence of these features in a patient suggested a strong prediction towards having the defined outcome (eFigure 3B and 3C).

**Calibration**

When considering dispensations predicted to be in the highest percentiles of risk, the top 5-percentile captured 42% of all outcomes using the XGBoost and logistic regression classifiers (Table 1). Also, as the predicted risk percentiles get higher (top 10 percentile to top 0.1 percentile), so too do the corresponding PPVs with the top 0.1 percentile associated with a PPV of 33% for the XGBoost classifier. As well, lower categories of risk percentiles were associated with lower outcomes (Figure 3, eFigure 4). When we simulated a monitoring workflow scenario with daily data uploads, a similar pattern was illustrated where the dispensations predicted to be higher risk had higher event rates (Figure 4).

After using the XGBoost and logistic regression classifiers to identify the dispensations in the highest predicted risk percentiles, the pre-test probability of the outcome (1.6%) was transformed into higher post-test probabilities, with higher probabilities in the riskier percentiles (Table 1). The number needed to screen also decreased as predicted risk increased (Table 1).

Comparing discrimination performance, ML risk prediction outperformed the current guideline approaches when using various combinations of guideline recommendations (Table 2). In many of the guideline scenarios, the estimated AUROCs were close to the 0.5 mark. When we estimated the discrimination performance of the logistic regression classifier based on database source, using all databases produced an AUROC of 0.88. Reducing the database source to only DAD, NACRS, Claims (co-morbidities only) resulted in an AUROC of 0.85, while PIN (prescription history) only was 0.78 (Table 3).

## Discussion

This study showed that ML techniques using available administrative data (prescription histories and ICD codes) may provide enough discriminatory performance to predict adverse outcomes associated with opioid prescribing. Indeed, our ML analyses showed very high discrimination performance at 0.88. The linear model (logistic regression) and XGBoosted Trees carried higher discrimination and calibration performance, while the neural network classifier did not perform as well. By identifying the predicted top 5-10 percentile of absolute risk pursuant to an opioid dispensation, we were able to capture approximately half of all outcomes using ML methods. All ML models we trained had higher discrimination performance using independent (external) validation sets than the clinical guideline approach.

Since the prevalence of our defined outcome is relatively low in the general population, PPVs would also be expectedly low. However, estimated PPVs increased when we considered higher risk dispensations, as is expected since PPV is related to event prevalence. This is important because different users of a risk predictor will require different predictive

capabilities. Similarly, our estimates of positive likelihood ratios and associated post-test probabilities also increased in dispensations with higher predicted risk indicating the strong predictive capabilities of the XGBoost and logistic regression classifiers; likelihood ratios >10 generate conclusive changes from pre-test to post-test probabilities[20].

The current guideline approach to assess absolute opioid prescribing risk produced c-statistic estimates closer to 0.5 indicating that discrimination was not much better than chance alone. ML models with higher predictive performance can better support health departments and PMPs with monitoring mandates to identify and intervene on those at high risk and their associated prescribers. We also found that adding co-morbidity features from administrative databases increased prediction performance compared to prescription history alone, thus making the case for the use of this data by PMPs and health departments. However, if only prescription history is available, our trained XGBoost classifier still had strong discrimination performance.

We found only one study that used ML approaches to quantify the absolute risk of an event pursuant to an opioid dispensation[10]. Their methodology used rolling 3-month windows for estimating risk and ML model training while we used historic records to estimate 30-day risk. Differences in study population and feature selection may explain why their highest performing ML model was deep learning (neural network classifier) and ours was not. Nevertheless, we were able to replicate their predictive performance using our ML approach as we both showed that ML approaches have higher predictive capabilities than guideline approaches. Both of our studies used predicted percentile risk estimates to identify high risk dispensations and were able to do so with strong discrimination and calibration performance.

Furthermore, we emphasized prognostic metrics which are more informative to assess the clinical utility of ML classifiers using pre- and post-test probabilities, something not done in other studies and recommended in medical guidelines[20]. This major aspect of our study, not done previously, is important because any ML classifier that does not increase prognostic information compared to baseline cannot be incorporated into decision making for the purpose of intervening on higher risk instead of lower risk patients. Indeed, another study we found describes how identifying cases in higher predicted risk percentiles using ML methods can be deployed in hospital settings for the purpose of targeted interventions[37] upon discharge, however the effect on outcomes is still to be determined.

The limitations of our study are similar to other ML studies[10] and need to be addressed when considering deployment of ML risk predictors. Our training dataset was not able to account for non-prescription opioid consumption and the risk associated with non-prescription use, both of which are substantial contributors to overall risk[2]. Regarding our analysis, we assumed that all dispensations were independent events; future research in this area should focus on employing ML methods using correlated data. As with all ML projects, our models were trained using Alberta data and might not be generalizable to other populations, or to specific populations within Alberta. However, our analyses were done on a large population and these results would be expected to be generalizable to the vast majority of patients. Moreover, one of the benefits of the ML process is that models can be retrained or similar methods could be used to develop new models to accommodate different populations.

This study suggests that ML risk prediction can support PMPs, especially if readily available administrative health data is used. PMPs currently use population-based guidelines

which we, and others, have shown cannot predict absolute individual risk. The ML process allows for model training, validation and deployment to specific settings in which, for the case of PMPs, high risk patients can be identified and targeted for intervention either at the patient or provider level. Moreover, ML classifiers can be retrained over time as changes in populations and trends in prescribing occur and are therefore specific to the population unlike broadly based guidelines. Further research can assess whether implementation of a ML-based monitoring system by PMPs leads to improved clinical outcomes.

## References


1. Belzak L, Halverson J. Evidence synthesis - The opioid crisis in Canada: a national perspective. *Health Promotion and Chronic Disease Prevention in Canada.* 2018;38(6):224-233.
2. Gomes T, Khuu W, Martins D, et al. Contributions of prescribed and non-prescribed opioids to opioid related deaths: population based cohort study in Ontario, Canada. *BMJ.* 2018;362:k3207.
3. Busse JW, Craigie S, Juurlink DN, et al. Guideline for opioid therapy and chronic noncancer pain. *Canadian Medical Association Journal.* 2017;189(18):E659-E666.
4. Dowell D. CDC guideline for prescribing opioids for chronic pain. 2016.
5. ismp Canada. Essential Clinical Skills for Opioid Prescribers. 2017; https://www.ismp-canada.org/download/OpioidStewardship/Opioid-Prescribing-Skills.pdf. Accessed Nov 2018.
6. Centre for Effective Practice. Management of Chronic Non Cancer Pain. 2017; thewellhealth.ca/cncp.
7. College of Physicians and Surgeons of Alberta. TPP Alberta – OME and DDD Conversion Factors. 2020; http://www.cpsa.ca/tpp/. Accessed Jun 2020.
8. World health Organization. Classification of Diseases (ICD). 2019; https://www.who.int/classifications/icd/icdonlineversions/en/. Accessed Jun 2020.
9. Gomes T, Mamdani MM, Dhalla IA, Paterson JM, Juurlink DN. Opioid Dose and Drug-Related Mortality in Patients With Nonmalignant PainOpioid Dose and Drug-related Mortality. *JAMA Internal Medicine.* 2011;171(7):686-691.
10. Lo-Ciganic W-H, Huang JL, Zhang HH, et al. Evaluation of Machine-Learning Algorithms for Predicting Opioid Overdose Risk Among Medicare Beneficiaries With Opioid Prescriptions. *JAMA network open.* 2019;2(3):e190968-e190968.
11. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. *JAMA.* 2017;318(14):1377-1384.
12. Shah NH, Milstein A, Bagley P, Steven C. Making Machine Learning Models Clinically Useful. *JAMA.* 2019;322(14):1351-1352.
13. Liu Y, Chen P-HC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA.* 2019;322(18):1806-1816.
14. Bastanlar Y, Ozuysal M. Introduction to machine learning. *Methods in molecular biology (Clifton, NJ).* 2014;1107:105-128.
15. Thottakkara P, Ozrazgat-Baslanti T, Hupf BB, et al. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PloS one.* 2016;11(5):e0155705.
16. Alberta Machine Intelligence Institute. Machine Learning Process Lifecycle. In:2019.
17. Hsich E, Gorodeski EZ, Blackstone EH, Ishwaran H, Lauer MS. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circulation: Cardiovascular Quality and Outcomes.* 2011;4(1):39-45.
18. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Paper presented at: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining2015.
19. Yusuf M, Atal I, Li J, et al. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. *BMJ open.* 2020;10(3):e034568.
20. Jaeschke R, Guyatt GH, Sackett DL, et al. Users' Guides to the Medical Literature: III. How to Use an Article About a Diagnostic Test B. What Are the Results and Will They Help Me in Caring for My Patients? *JAMA.* 1994;271(9):703-707.
21. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA.* 2018;320(21):2199-2200.

22. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine.* 2015;162(1):W1-W73.
23. equator network. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. 2020; https://www.equator-network.org/reporting-guidelines/tripod-statement/. Accessed Feb 2020.
24. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open.* 2016;6(11):e012799.
25. Buitinck L, Louppe G, Blondel M, et al. API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:13090238.* 2013.
26. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Paper presented at: Advances in neural information processing systems2017.
27. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Paper presented at: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining2016.
28. The pandas development team. pandas-dev/pandas: Pandas. 2020; https://doi.org/10.5281/zenodo.3509134, Jan 2021.
29. Zhou H, Della PR, Roberts P, Goh L, Dhaliwal SS. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. *BMJ Open.* 2016;6(6):e011060.
30. Brownlee J. A Gentle Introduction to Imbalanced Classification. 2020; https://machinelearningmastery.com/what-is-imbalanced-classification/. Accessed Jan 2021.
31. Government of Canada. Forward Sortation Area—Definition. 2015; https://www.ic.gc.ca/eic/site/bsf-osb.nsf/eng/br03396.html. Accessed April 2020, 2020.
32. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical care.* 2005:1130-1139.
33. World Health Organization. International language for drug utilization research, ATC/DDD. 2020; https://www.whocc.no/. Accessed Jun 2020, 2020.
34. College of Physicians and Surgeons of Alberta. OME and DDD conversion factors. http://www.cpsa.ca/wp-content/uploads/2017/06/OME-and-DDD-Conversion-Factors.pdf.
35. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European heart journal.* 2017;38(23):1805-1814.
36. Rose S. Machine Learning for Prediction in Electronic Health Data. *JAMA Network Open.* 2018;1(4):e181404-e181404.
37. Morgan DJ, Bame B, Zimand P, et al. Assessment of Machine Learning vs Standard Prediction Rules for Predicting Hospital Readmissions. *JAMA Network Open.* 2019;2(3):e190348-e190348.
38. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE.* 2015;10(3):e0118432.
39. Shah ND, Steyerberg EW, Kent DM. Big Data and Predictive Analytics: Recalibrating Expectations. *JAMA.* 2018;320(1):27-28.
40. Molnar C. *Interpretable machine learning. A Guide for Making Black Box Models Explainable.* 2019.
41. Centers for Medicare & Medicaid Services (CMS). Announcement of calendar year (CY) 2019 Medicare Advantage capitation rates and Medicare Advantage and Part D payment policies and final call letter.

## Figure Legend

Figure 1. Patient flow diagram of study participants used for training and validating ML models. NACRS: National Ambulatory Care Reporting System; DAD: Discharge Abstract Database; VS: Vital Statistics; PIN: Pharmaceutical Information Network; Claims: Physician Claims

Figure 2. Area under the receiver operating characteristic curve (AUROC) (A) and precision-recall curves (B) for all dispensations using logistic regression (L1), neural network, support vector machine (SVM), XGBoost and Naïve-Bayes; precision-recall curves for higher risk dispensations according to predicted risk percentile categories for logistic regression (C) and XGBoost (D) using the 2018 validation set.

Figure 3. Calibration curves plotting observed vs. quantiles of estimated risk for XGBoost (A) and logistic regression (B) classifiers using the 2018 validation dataset. For both classifiers, the majority of counts (dispensations) were predicted to be lower risk.

Figure 4. Simulation of a clinical workflow with daily uploads and events per 100 daily dispenses by risk percentiles using 2018 Quarter 1 (Q1) data for logistic regression (A) and XGBoost (B) classifiers. For both classifiers, the majority of counts (dispensations) were predicted to be lower risk.

Table 1. Highest percentiles of estimated risk and predictive performance using the XGBoost and logistic regression classifiers for the 2018 validation dataset (n=393,023). Total number of dispenses= 1,977,389; total number of outcomes= 31,392.

| Metric | Top 0.1%ile | | Top 1%ile | | Top 5%ile | | Top 10%ile | |
|---|---|---|---|---|---|---|---|---|
| | XGBoost | Logistic Regression | XGBoost | Logistic Regression | XGBoost | Logistic Regression | XGBoost | Logistic Regression |
| **Number of Dispenses** | 1,977 | 1,977 | 19,774 | 19,774 | 98,869 | 98,869 | 197,739 | 197,739 |
| **TP captured** | 655 | 472 | 4204 | 4100 | 13224 | 13293 | 18404 | 18409 |
| **Percent of TP** | 2.09 | 1.50 | 13.39 | 13.06 | 42.13 | 42.35 | 58.63 | 58.64 |
| **FP captured** | 1322 | 1505 | 15570 | 15674 | 85645 | 85576 | 179335 | 179330 |
| **PPV** | 33.13 | 23.87 | 21.26 | 20.73 | 13.38 | 13.45 | 9.31 | 9.31 |
| **PLR** | 30.71 | 19.44 | 16.74 | 16.22 | 9.57 | 9.63 | 6.36 | 6.36 |
| **Post-test Probability*** | 33.13 | 23.87 | 21.26 | 20.73 | 13.38 | 13.45 | 9.31 | 9.31 |
| **NNS** | 3.17 | 4.49 | 5.08 | 5.22 | 8.48 | 8.43 | 12.95 | 12.95 |

*Pre-test probability estimated at 1.6% using prevalence.

TP: true positives; FP: false positives; PPV: positive predictive value; PLR: positive likelihood ratio; NNS: number needed to screen

Note: Logistic regression used L1 (lasso) parameter regularization

Table 2. Discrimination performance of guideline approach using the 2018 validation set. Guideline approaches were adapted from the 2017 Canadian Opioid Prescribing Guideline and 2019 Centers for Medicare & Medicaid Services (CMS) opioid safety measures and compared to logistic regression and XGBoost classifiers (each with an estimated area under the receiver operating characteristic curve of 0.88).

| **Canadian Guidelines *** | **AUROC** | **Sensitivity** | **Specificity** |
|---|---|---|---|
| History of mental disorder only | 0.620 | 0.90 | 0.34 |
| Substance abuse only | 0.686 | 0.99 | 0.37 |
| OME/day >90 only | 0.539 | 0.22 | 0.85 |
| (Mental disorder and substance abuse) OR OME/day >90 | 0.690 | 0.91 | 0.47 |
| Mental disorder and substance abuse AND OME/day >90 | 0.560 | 0.20 | 0.91 |
| Mental disorder OR substance abuse OR OME/day >90 | 0.589 | 0.99 | 0.18 |
| **CMS Guidelines**** | | | |
| High opioid dose (>120 OME/day for 90+days) | 0.507 | 0.081 | 0.933 |
| Concurrency (Opioid & BZRA for 30+ days) | 0.575 | 0.423 | 0.727 |
| Multiple doctors (>4) | 0.591 | 0.294 | 0.888 |
| Multiple pharmacies (>4) | 0.537 | 0.120 | 0.959 |
| All conditions | 0.50 | 0.001 | 0.999 |
| Any condition | 0.622 | 0.62 | 0.625 |

OME: daily oral morphine equivalents; BZRA: benzodiazepine receptor agonist. Elixhauser scoring ICD codes were used to identify mental disorders and substance abuse.

*The Canadian guidelines do not specify timelines. >90 OME was determined by taking the average daily OME over the 30 days prior to dispensation

**The CMS guidelines specify a timeline of 90 or more days at >120 OME and concurrent use of opioids and benzodiazepines for 30 days or more

Table 3. Discrimination performance based on database source using area under the receiver operating characteristic curve (AUROC) for the logistic regression classifier on the 2018 validation set.

| Database source | Predictor Variables formed from database | AUROC |
|---|---|---|
| PIN only | Drug utilization + Prescription history (ATC level 3) | 0.78 |
| DAD, NACRS, Claims | Co-morbidities | 0.85 |
| PIN, DAD NACRS, Claims (all databases used in study) | Demographic + Drug Utilization + Healthcare Utilization + Co-morbidities | 0.88 |

Note: drug utilization includes features describing oral morphine equivalents[34], concurrent use with benzodiazepines, number of opioid and benzodiazepine dispensations, number of unique opioid and benzodiazepine molecules; health care utilization includes features describing number of unique health providers visited, number of hospital visits; logistic regression used L1 (lasso) parameter regularization

Figure 1. Patient flow diagram of study participants used for training and validating ML models. NACRS: National Ambulatory Care Reporting System; DAD: Discharge Abstract Database; VS: Vital Statistics; PIN: Pharmaceutical Information Network; Claims: Physician Claims

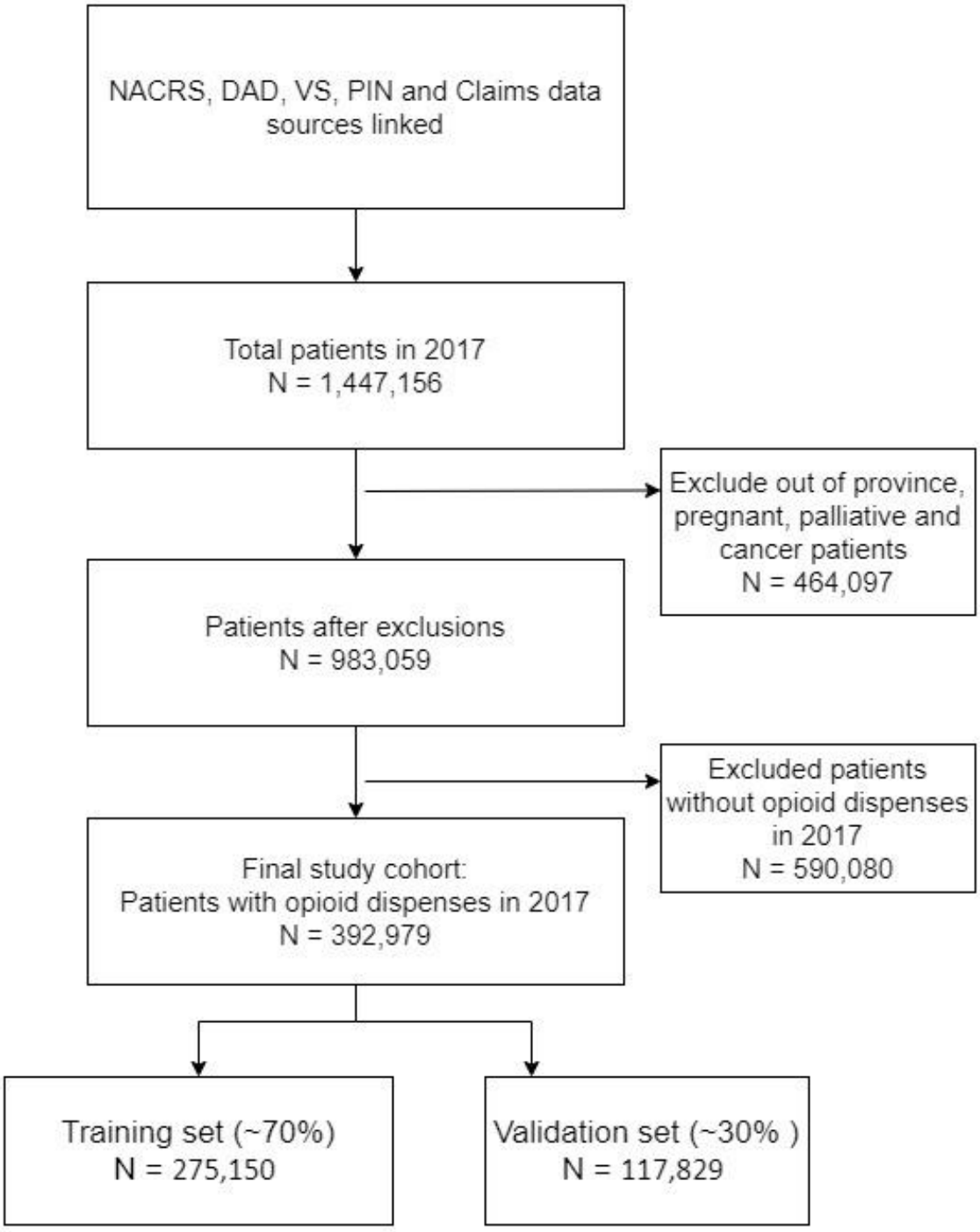
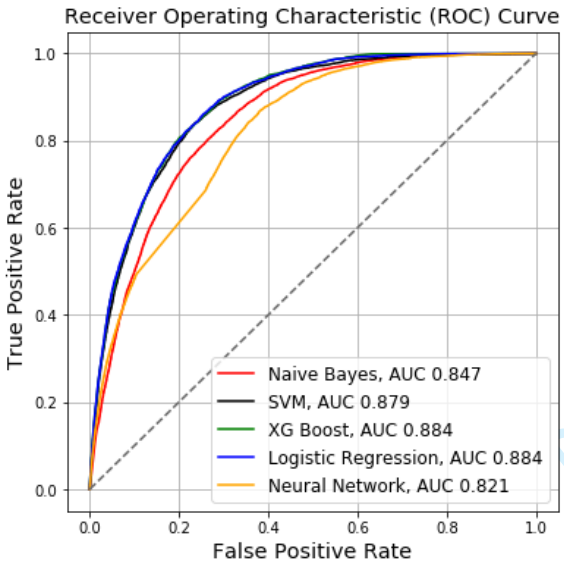
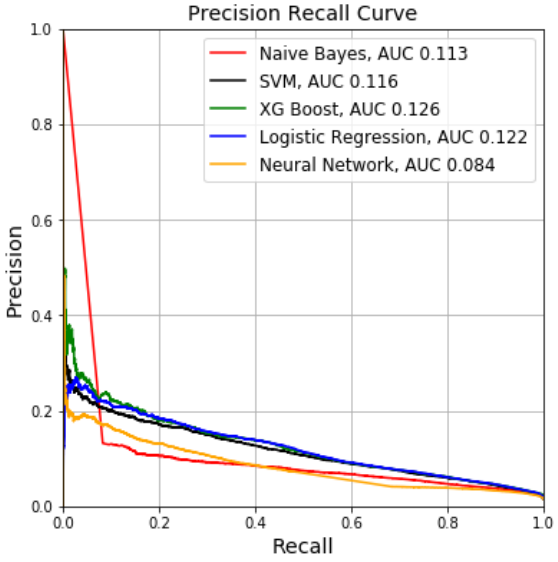


Figure 2. Area under the receiver operating characteristic curve (A) and precision-recall curves (B) for all dispensations using logistic regression (L1), neural network, support vector machine (SVM), XGBoost and Naïve-Bayes; precision-recall curves for higher risk dispensations according to predicted risk percentile categories for logistic regression (C) and XGBoost (D) using the 2018 validation set.

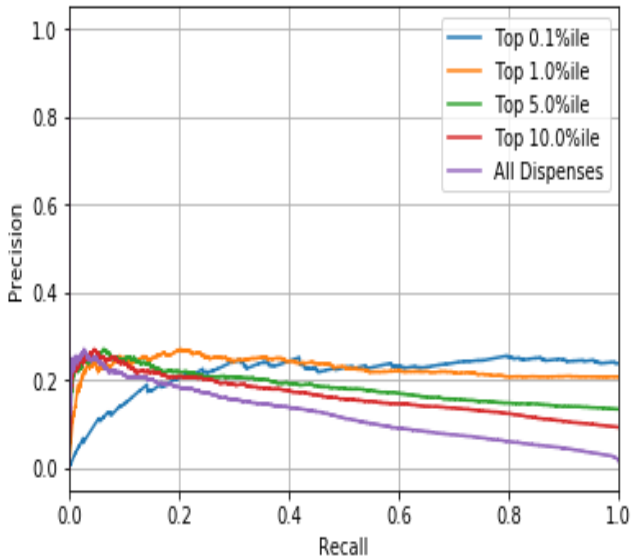
(A)



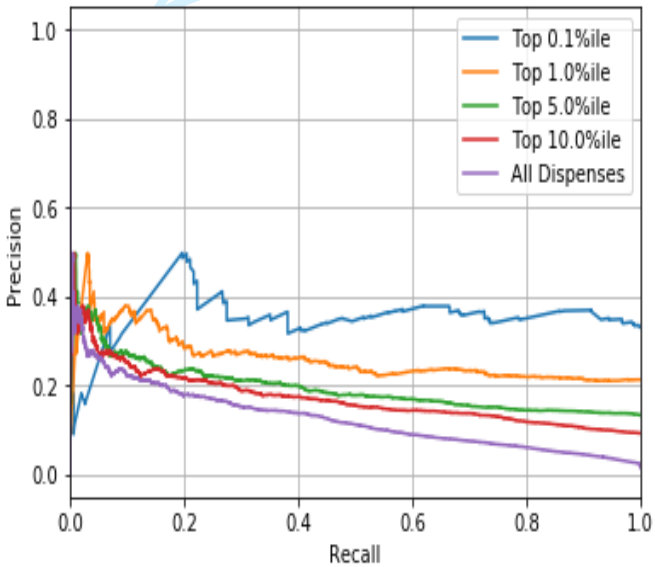

(B)




(C) Logistic Regression




(D) XGBoost




AUC: area under the curve

Figure 3. Calibration curve plotting observed vs. quantiles of estimated risk for the XGBoost classifier using the 2018 validation dataset. The majority of counts (dispensations) were predicted to be lower risk.

(A) XGBoost

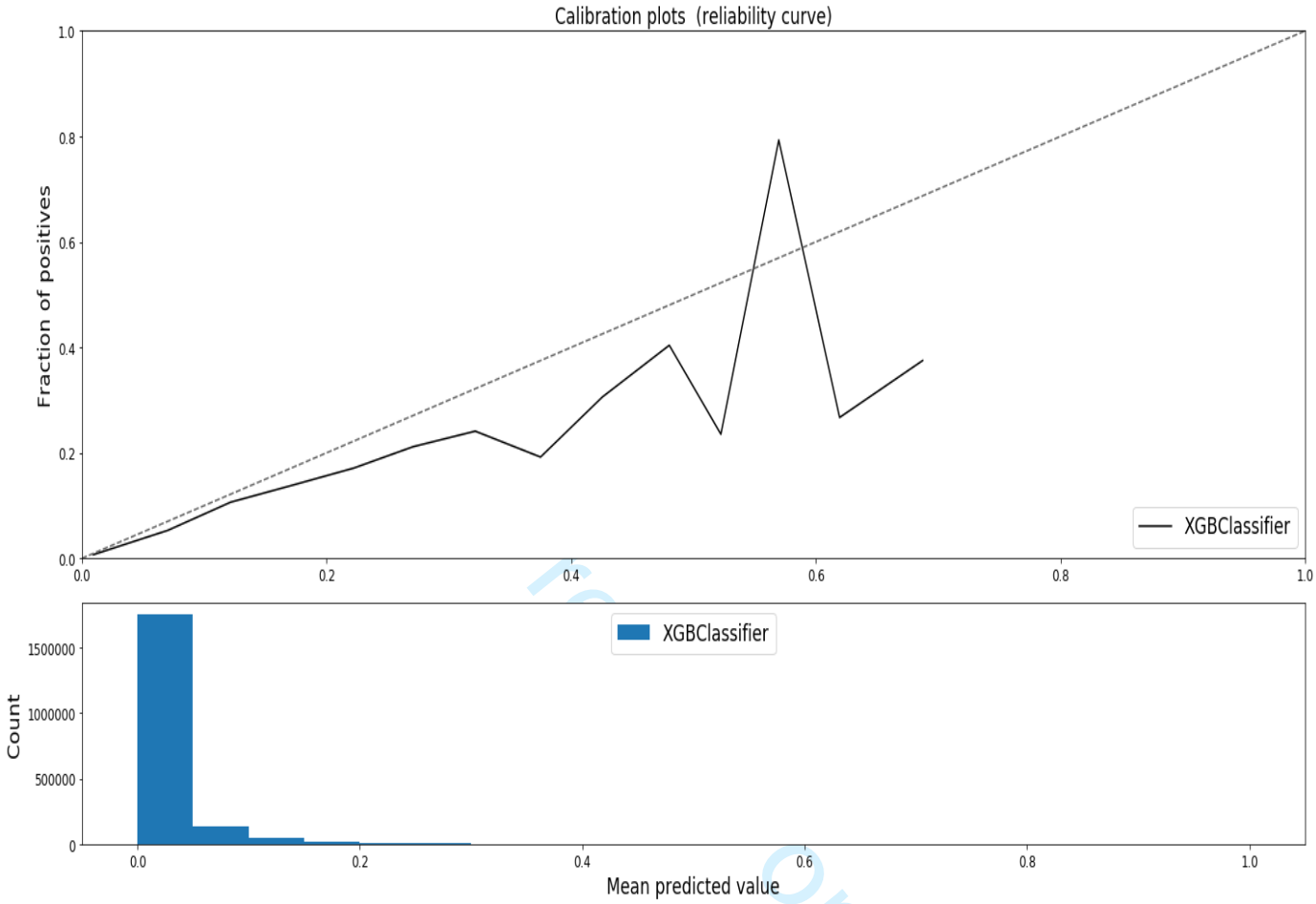
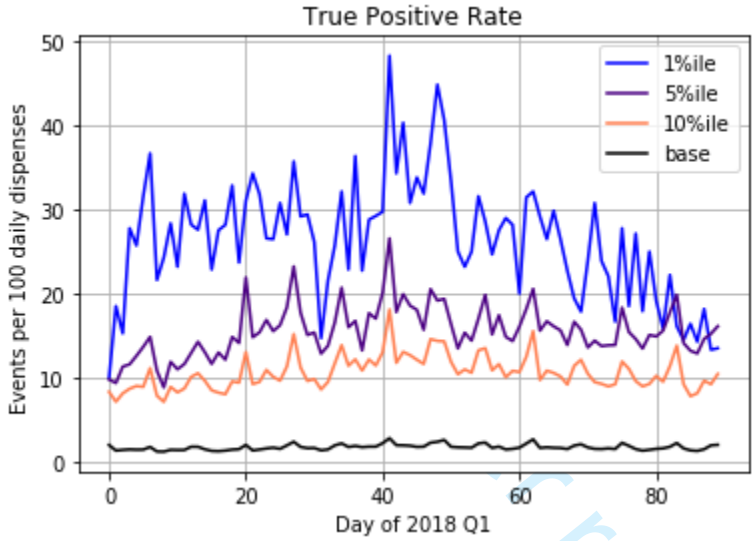
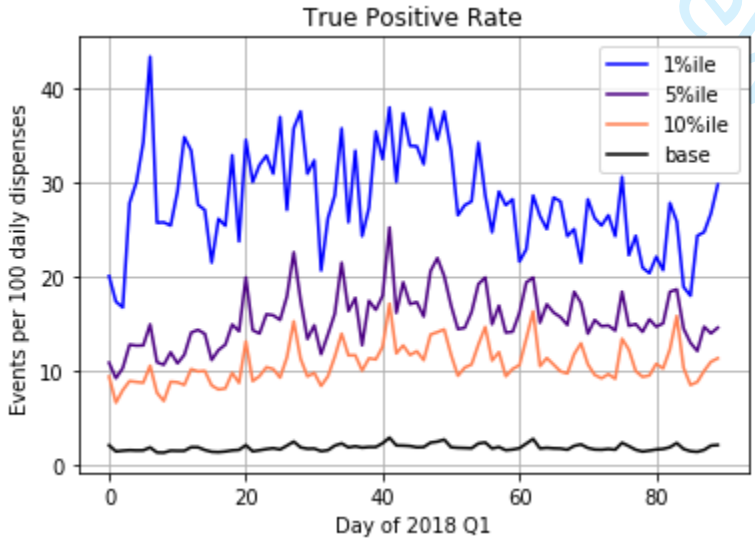


Figure 4. Simulation of a clinical workflow with daily uploads and events per 100 daily dispenses by risk percentiles using 2018 Quarter 1 (Q1) data for logistic regression (A) and XGBoost (B) classifiers. For both classifiers, the majority of counts (dispensations) were predicted to be lower risk.

(A) Logistic Regression (L1)




(B) XGBoost



**Supplementary Content**

**eAppendix.** Machine learning algorithms

**eTable 1.** Diagnostic codes used to exclude patients who had cancer, were pregnant, or were under palliative care.

**eTable 2.** Diagnostic codes used to identify the defined study outcome from emergency visit, hospitalization and death data.

**eTable 3.** Baseline characteristics of study patients (n=392,979). Co-morbidities were determined using Elixhauser criteria. All p-values in the $chi^2$ test of independence were <0.001 unless otherwise indicated.

**eTable 4.** Characteristics of study participants between training and validation groups using 2017 data.

**eTable 5.** Candidate predictors used to train ML algorithms.

**eTable 6.** Discrimination performance using area under the receiver operating characteristic curve (AUROC) of various ML algorithms. Training and validation were done using 2017 data (n=393,979); another independent validation was performed using 2018 data (n=393,023).

**eFigure1.** Schematic of study design and feature generation

**eFigure2**. Feature importance from logistic regression and tree-based (XGBoost) classifiers using the 2018 validation set.

**eFigure3**. Shapley values and feature impact in the XGBoost classifier using the 2018 validation set to describe "associations" between features and the outcome.

**eFigure 4.** Calibration curve plotting observed vs. quantiles of estimated risk for the logistic regression (L1) classifier using the 2018 validation dataset. The majority of counts (dispensations) were predicted to be lower risk.

**eReferences.**

eAppendix. Machine Learning Algorithms

**Introduction**

While there are always updates and new methods coming up in the fields of machine learning, in this study, we have focused on some of the most reliable and proven approaches for predictive modelling which are explainable and popularly used in previous studies of similar nature.

**Logistic Regression**

Regression analysis models the relationship between a dependent variable and a set of independent variables [1]. Typically, this includes understanding how the value of the dependent variable changes with the changes in the values of independent variables. Logistic regression [1] uses the logistic function to model a binary dependent variable, where, based on the values of the independent variables the model can approximate one of the two classes, the instance belongs to. This basic binary model can be extended to deal with multiple classes (e.g. One-vs-all classifiers). However, logistic regression is only capable of modeling a linear relationship of independent variables to the dependent variable, hence limited to problems with linear decision boundaries. We used the sci-kit learn library in our experiments[6] and found L1 regularization to be more effective.

**Ridge Classifier**

We used the ridge classifier implemented in the Scikit learn library[5]. It implements a classifier using ridge regression which uses an L2 regularization on the least square objective function. The library converts the labels into -1 and 1 and fits a linear regression on the converted labels with the regularization.

**Random Forest**

Random forest is a tree ensemble learning algorithm that has wide applicability in many domains[1]. Random forest is a nonlinear learning algorithm, which arrives at nonlinear decision boundaries by independently combining multiple decision trees. Each individual decision tree in the forest can be grown independently of each other on a subset of the training data. Random forests are mainly sensitive to the number of trees, the depth of a tree and the number of covariates randomly chosen to split at each node[1]. These hyper-parameters can be tuned to find the best configuration of every dataset. Random Forests, in general, are less prone to overfit since they always grow individual trees on a subset of the training data[1]. At prediction time, the decision of each tree is aggregated to compute the final prediction.

**Neural Networks (NN)**

Neural networks are another collection of non-linear learning algorithms with high representation power. They are known to be able to find mappings from an input to an output from a larger non-linear function space [2]. This ability to represent a larger space of nonlinear

functions has shown to be very effective recently in many application domains such as natural language processing, computer vision, genomics, computer games and health[2]. Neural networks come in many flavors learning nonlinear mapping of different types of data such as Convolutional NNs being most effective with images and Recurrent NNs for time series and language data. Identifying the most effective neural network structure is one of the difficult and the most time-consuming aspect of applying neural networks to new application domains and data. Generally, neural networks try to exploit the relationships in the raw unstructured data (eg: image and text) presented to the network but with more structured data such as health records and ICD codes learning relationships is much complex. Our neural network models are mainly based on densely connected hidden layers with ReLu[6] activation function. We used the cross-entropy loss for the binary classification Adam optimizer. We used a simple feed forward network using Sklearn MLP classifier with hyperparameter tuning for the NN.

**Boosted Learning Algorithms**

Boosting is a process to ensemble multiple base learning algorithms to arrive at better overall performance than any individual base learner[1]. In contrast to independently building multiple models from the subsets of the data, boosting re-weights the training data every time a model is learned for future models. This weighting happens to give more preference to currently misclassified data points in the next round compared to the correctly classified data points. Therefore future learners try to do better on the misclassified data points leading to a collection base learners having a better-combined prediction. This process is sequential so each base learner is dependent on the output of the previously trained model (it is worthy to note XGBoost provides a parallel tree boosting alternative). In our work, we have experimented with several boosting meta-learning algorithms such as XGBoost[7], AdaBoost[5] and GBM[5]. XGBoost uses a variant of trees as the base learner whereas AdaBoost (from Sci-kit learn) can use many ML algorithms as base learners. GBM uses logistic regression by default as the base learner. We used all 3 types of boosting with tuned hyperparameters for comparison.

**Naive Bayes**

Naive Bayes is based on the Bayes theorem with a strong independence assumption between the covariates[1]. This assumption helps in building a simple probabilistic model for learning and inference. Naive Bayes coefficients scale linearly with the number of covariates making this a suitable model for high-dimensional data. We used Naive Bayes as a simple baseline learning algorithm for comparison.

**Support Vector Machines (SVM)**

SVMs[4] are maximum margin classifiers optimizing for learning a hyperplane having the maximum distance away from each of the class data points[1]. SVM is a linear classifier but with the kernel trick to map the inputs to the higher dimensional space, it can learn nonlinear decision boundaries in the input space. SVMs are very effective binary classifiers with the kernel trick[1]. With larger datasets, SVMs tend to become more computationally intensive.

eTable 1. Diagnostic codes used to exclude patients who had cancer, were pregnant, or were under palliative care.

| Condition | ICD 9 | ICD 10 |
|---|---|---|
| **Cancer** | 140.x - 239.x | C00.x - C99.x, D00.x - D49.x |
| **Pregnancy** | 630.x - 679.x | O00.x - O99.x |
| **Palliative** | V66 | Z51.0, Z51.1, Z51.5 |

eTable 2. Diagnostic codes used to identify the defined study outcome from emergency visit, hospitalization and death data.

| ICD 10 | Condition |
|---|---|
| **T40.x** | Poisoning by, adverse effect of and underdosing of narcotics and psychodysleptics |
| **F55.x** | Abuse of non-psychoactive substances |
| **F11.x - F19.x** | Mental and behavioral disorders due to psychoactive substance use |

eTable 3. Baseline characteristics of study patients (n=392,979). Co-morbidities were determined using Elixhauser criteria. All p-values in the chi$^2$ test of independence were <0.001 unless otherwise indicated.

| Characteristic | Number without Event n=386,371 | Percent | Number with Event n=6,608 | Percent |
|---|---|---|---|---|
| **Age:** | | | | |
| Mean (SD) | 48.1 (16.4) | -- | 41.2 (12.4) | -- |
| 18-45 | 162057 | 41.9 | 3466 | 52.4 |
| 45-65 | 154632 | 40.0 | 2656 | 40.2 |
| >65* | 69682 | 18.0 | 486 | 7.4 |
| **Male** | 197491 | 50.3 | 3922 | 59.4 |
| **Female** | 194794 | 49.7 | 2686 | 40.6 |
| **Alcohol Disorder** | 66320 | 16.9 | 5220 | 79.0 |
| **Arrhythmia** | 90621 | 23.1 | 1959 | 29.6 |
| **Blood Loss Anemia** | 1164 | 0.3 | 82 | 1.2 |
| **Congestive Heart Failure** | 18954 | 4.8 | 565 | 8.6 |
| **Coagulopathy** | 8053 | 2.1 | 356 | 5.4 |
| **Deficiency Anemia** | 34188 | 8.7 | 971 | 14.7 |
| **Depression** | 159140 | 40.6 | 5518 | 83.5 |
| **Diabetes**** | 64132 | 16.3 | 1408 | 21.3 |
| **Substance Abuse Disorder** | 74678 | 19.0 | 5485 | 83.0 |
| **Fluid Disorder** | 42690 | 10.9 | 3012 | 45.6 |
| **Hypertension**** | 140171 | 35.7 | 2624 | 39.7 |
| **Hypothyroidism** | 45519 | 11.6 | 601 | 9.1 |
| **Injury^** | 195688 | 49.9 | 5541 | 83.9 |
| **Liver Disorder** | 21656 | 5.5 | 1588 | 24.0 |
| **Neurologic Disorder** | 230490 | 58.8 | 5387 | 81.5 |
| **Obesity** | 63393 | 16.2 | 970 | 14.7 |
| **Poisoning^** | 17434 | 4.4 | 2775 | 42.0 |
| **Psychoses** | 35870 | 9.1 | 3162 | 47.9 |
| **Renal Disorder** | 16166 | 4.1 | 499 | 7.6 |
| **Rheumatoid Conditions** | 111458 | 28.4 | 3157 | 47.8 |
| **HIV Infection** | 1098 | 0.3 | 141 | 2.1 |
| **Paralysis** | 3874 | 1.0 | 187 | 2.8 |
| **Peptic Ulcer Disease** | 11728 | 3.0 | 509 | 7.7 |
| **Pulmonary Circulation Disorder** | 9611 | 2.4 | 430 | 6.5 |
| **Chronic Pulmonary Disease** | 102990 | 26.3 | 2913 | 44.1 |
| **Peripheral Vascular Disease** | 14467 | 3.7 | 389 | 5.9 |
| **Valvular Disease** | 7308 | 1.9 | 226 | 3.4 |
| **Weight Loss** | 16207 | 4.1 | 747 | 11.3 |

*p-value for age >65 is an estimated 0.037

^ Injury: ICD10: S00-T98; Poisoning: ICD10: T36-T50

** Complicated, uncomplicated diabetes and hypertension were collapsed into one category each

eTable 4. Characteristics of study participants between training and validation groups using 2017 data.

| Characteristic | Number in training group N=275,150~ | Percent | Number in validation group N=117,829~ | Percent |
|---|---|---|---|---|
| **Age:** | | | | |
| **Mean (SD)** | 48.3 (16) | -- | 48.2 (16) | -- |
| **18-45** | 114356 | 41.5 | 49909 | 42.3 |
| **45-65** | 111859 | 40.7 | 47132 | 40.0 |
| **>65** | 48935 | 17.8 | 20788 | 17.6 |
| **Male** | 138603 | 48.5 | 59339 | 48.4 |
| **Female** | 136545 | 47.8 | 58490 | 47.7 |
| **Alcohol Disorder** | 46792 | 16.4 | 20199 | 16.5 |
| **Arrhythmia** | 63637 | 22.3 | 27201 | 22.2 |
| **Blood Loss Anemia** | 839 | 0.3 | 336 | 0.3 |
| **Congestive Heart Failure** | 13320 | 4.7 | 5694 | 4.6 |
| **Coagulopathy** | 5697 | 2.0 | 2393 | 2.0 |
| **Deficiency Anemia** | 24096 | 8.4 | 10179 | 8.3 |
| **Depression** | 112080 | 39.2 | 47628 | 38.9 |
| **Diabetes**** | 45131 | 15.8 | 19144 | 15.6 |
| **Substance Abuse Disorder** | 52609 | 18.4 | 22713 | 18.5 |
| **Fluid Disorder** | 30272 | 10.6 | 12780 | 10.4 |
| **Hypertension**** | 98546 | 34.5 | 41840 | 34.1 |
| **Hypothyroidism** | 31908 | 11.2 | 13666 | 11.2 |
| **Injury*** | 137423 | 48.1 | 58865 | 48.0 |
| **Liver Disorder** | 15252 | 5.3 | 6567 | 5.4 |
| **Neurologic Disorder** | 161706 | 56.5 | 69341 | 56.6 |
| **Obesity** | 44607 | 15.6 | 18882 | 15.4 |
| **Poisoning*** | 12503 | 4.4 | 5293 | 4.3 |
| **Psychoses** | 25422 | 8.9 | 10860 | 8.9 |
| **Renal Disorder** | 11403 | 4.0 | 4817 | 3.9 |
| **Rheumatoid Conditions** | 78268 | 27.4 | 33420 | 27.3 |
| **HIV Infection** | 774 | 0.3 | 336 | 0.3 |
| **Paralysis** | 2717 | 1.0 | 1176 | 1.0 |
| **Peptic Ulcer Disease** | 8239 | 2.9 | 3533 | 2.9 |
| **Pulmonary Circulation Disorder** | 6771 | 2.4 | 2877 | 2.3 |
| **Chronic Pulmonary Disease** | 72265 | 25.3 | 30949 | 25.3 |

| | | | | |
|---|---|---|---|---|
| **Peripheral Vascular Disease** | 10228 | 3.6 | 4278 | 3.5 |
| **Valvular Disease** | 5111 | 1.8 | 2215 | 1.8 |
| **Weight Loss** | 11477 | 4.0 | 4790 | 3.9 |

~p-values for chi$^2$ test of independence were all >0.06 when comparing training and validation sets.

*Injury: ICD10: S00-T98; Poisoning: ICD10: T36-T50

** Complicated, uncomplicated diabetes and hypertension were collapsed into one category each

eTable 5. Candidate predictors used to train ML algorithms.

| **Category (data source)** | **Description** |
|---|---|
| Demographic information (PIN) | age, sex, postal codes, mean income |
| Drug utilization history (PIN) | drug dispenses in past 30 days using on ATC codes, oral morphine equivalents, concurrent use with benzodiazepines defined as at least 7 days of cumulative concurrent use in the 30 days prior to dispensation, number of dispensations and unique molecules of opioids and benzodiazepines |
| Health care utilization (PIN DAD) | flags for previous hospitalizations, number of unique providers |
| ICD based co-morbidities (DAD, NACRS, Claims) | Elixhauser condition flags based on the past 5 years of claims, hospitalizations, and emergency visits. |

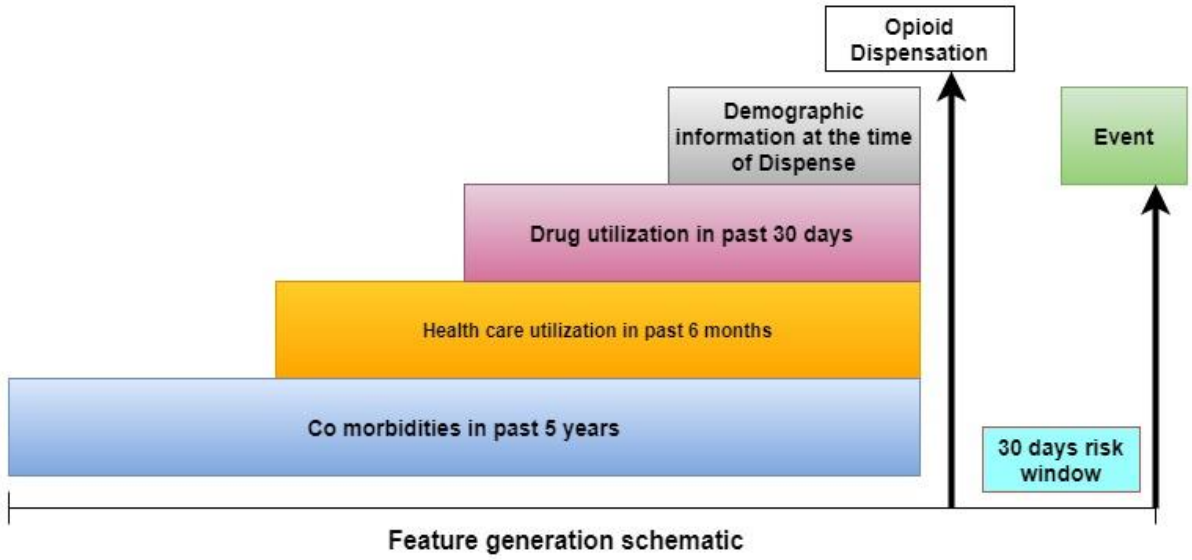
Note: ICD: International Statistical Classification of Diseases and Related Health Problems, World Health Organization.

eTable 6. Discrimination performance using area under the receiver operating characteristic curve (AUROC) of various ML algorithms using all features (demographics, health utilization, prescription history, co-morbidities). Training and validation were done using 2017 data (n=393,979); another independent validation was performed using 2018 data (n=393,023).

| Algorithm | Train | Validation 2017 | Validation 2018 |
|---|---|---|---|
| XGBoost Classifier | 0.897 | 0.870 | 0.884 |
| Logistic Regression | 0.887 | 0.869 | 0.884 |
| Gradient Boosting Classifier | 0.898 | 0.868 | 0.883 |
| AdaBoost Classifier | 0.884 | 0.868 | 0.882 |
| Random Forest Classifier | 0.909 | 0.863 | 0.881 |
| Ridge Classifier | 0.895 | 0.863 | 0.879 |
| SVM | 0.896 | 0.860 | 0.878 |
| Gaussian Naive Bayes | 0.846 | 0.826 | 0.847 |
| Decision Tree Classifier | 0.919 | 0.791 | 0.822 |
| Neural Networks | 0.827 | 0.804 | 0.821 |

Note: Logistic regression used L1 (lasso) parameter regularization

**eFigure 1. Schematic of study design and feature generation**



**eFigure2. Feature importance from logistic regression and tree-based (XGBoost) classifiers using the 2018 validation set. The logistic regression classifier relied more on co-morbidity data from DAD, NACRS, and Claims databases; XGBoost classifier relied more on data from the PIN database. AUROCs for both classifiers were similar at 0.88.**

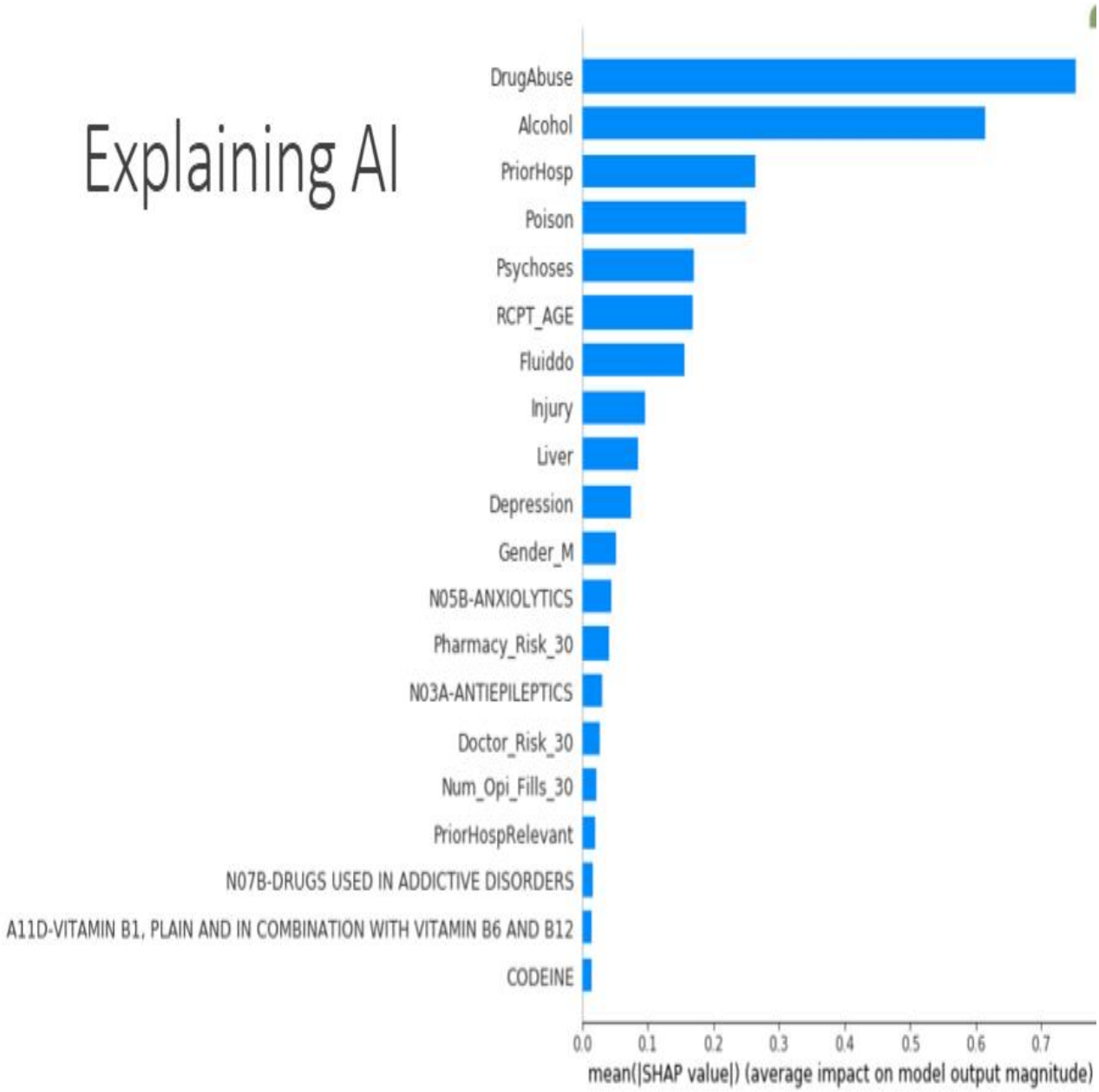
| LOGISTIC REGRESSION | |
|---|---|
| Drug Abuse | 1.00 |
| Age | 0.65 |
| Prior Hospitalization | 0.62 |
| Alcohol Abuse | 0.62 |
| Fluid Disorder | 0.32 |
| Substance Poison | 0.31 |
| Psychoses | 0.31 |
| Num_Benzo_Ingred_30 | 0.26 |
| Depression | 0.19 |
| Concurrent_Opioid_Benzo_30 | 0.19 |
| Injury | 0.17 |

| TREE-BASED MODEL | |
|---|---|
| Age | 1.00 |
| Num_Fills_30 | 1.00 |
| Num_Opioid_Fills_30 | 0.86 |
| Num_Benzo_Fills_30 | 0.46 |
| Doctor_Risk_30 | 0.45 |
| Total_OME_30_Days_Supply | 0.43 |
| Substance poison | 0.37 |
| Pharmacy_Risk_30 | 0.35 |
| Num_Doctors_30 | 0.34 |
| Income | 0.34 |
| Prior hospitalization | 0.26 |

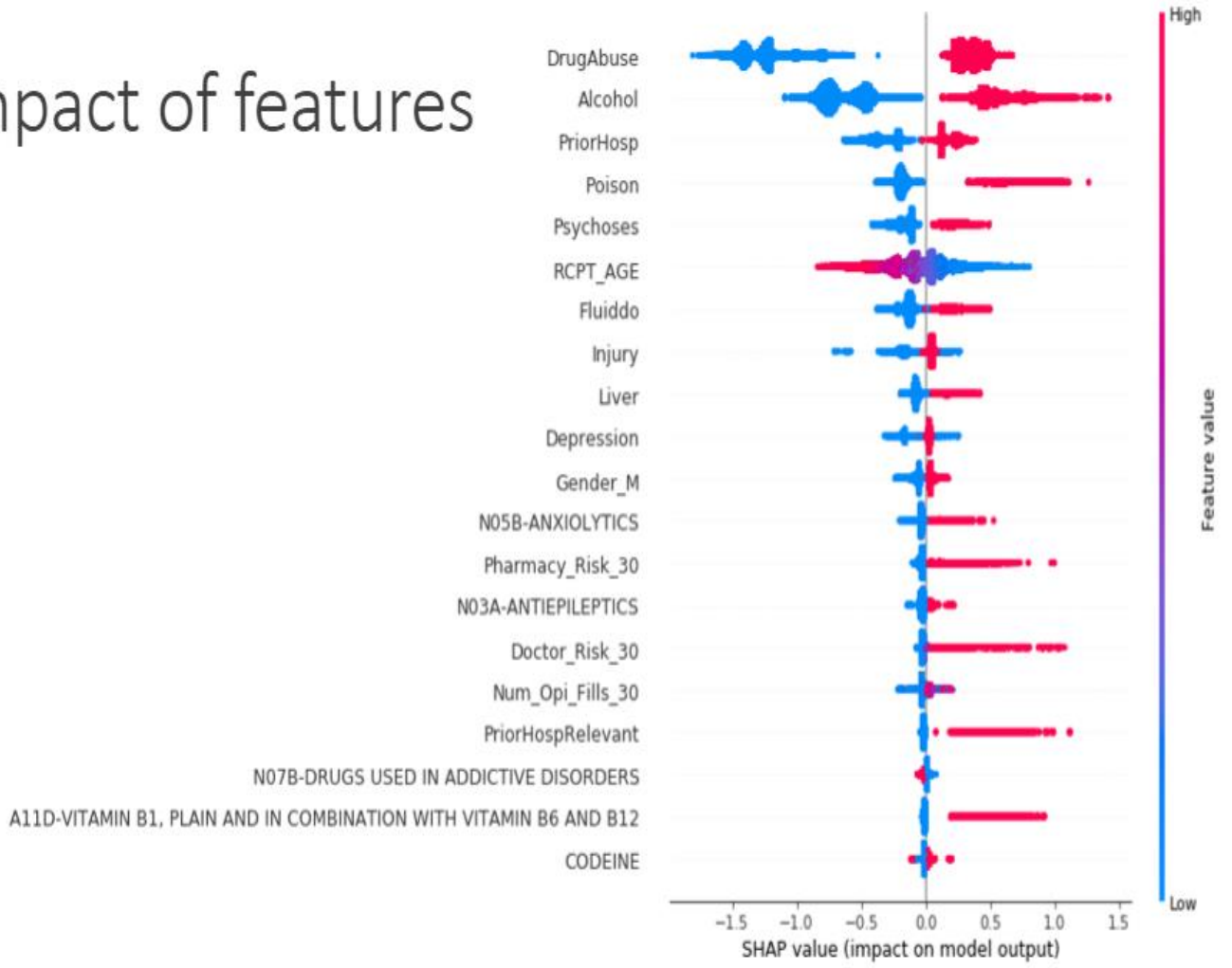
Note: Logistic regression used L1 (lasso) parameter regularization

**eFigure 3. Shapley values and feature impact in the XGBoost classifier using the 2018 validation set to describe "associations" between features and the outcome. Features with the most impact on the model with drug abuse with drug abuse ranked highest (A); tornado plot illustrating feature impact. Red indicates higher impact and plots to the right of 0.0 indicate the tendency to be associated with the study outcome while blue indicates lower impact and plots to the left of 0.0 indicate the tendency to be associated with no outcome (B); explaining the prediction of study outcomes based on predictor values for 4 patients (C).**

**(A)**



**(B)**

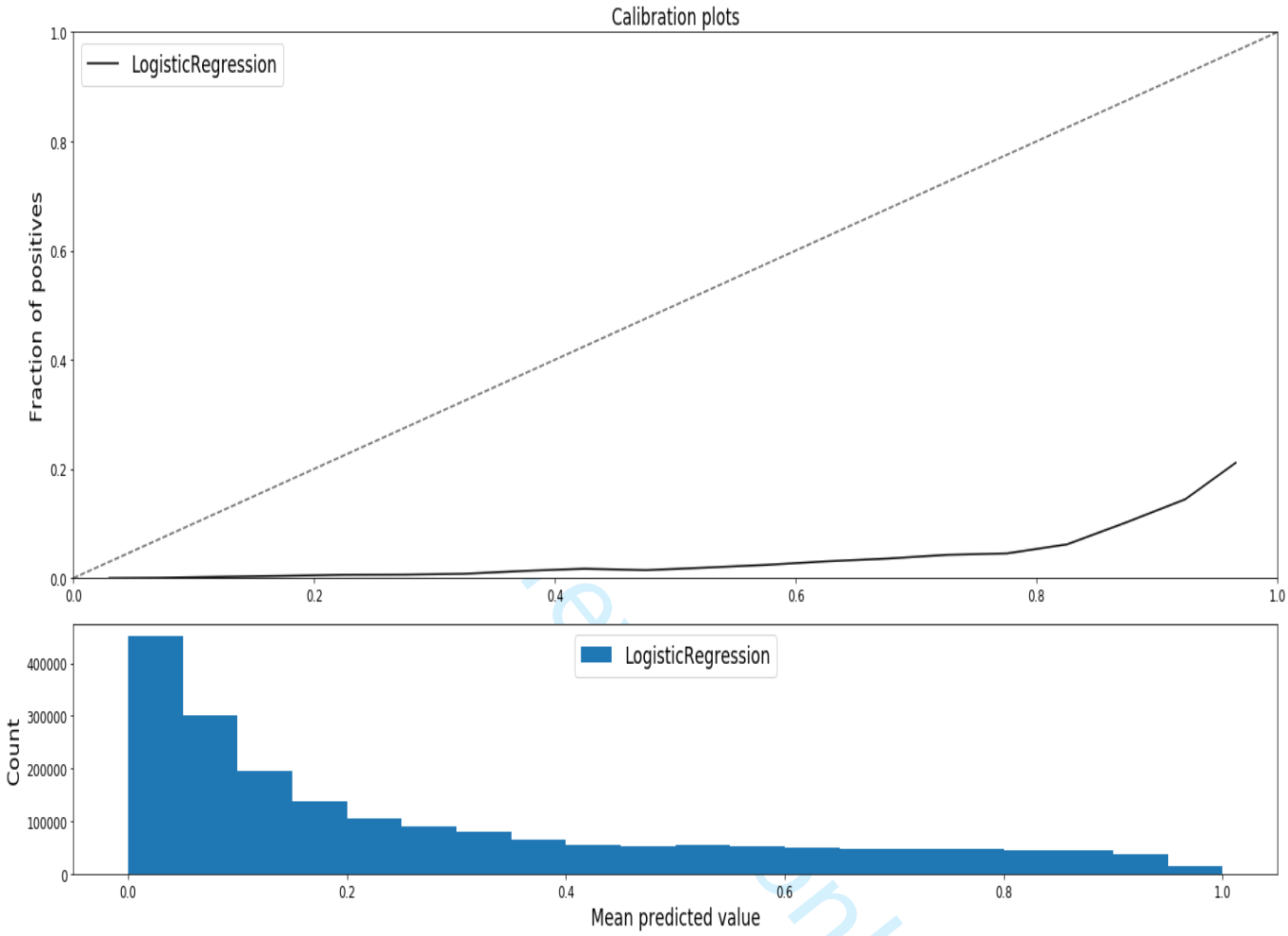



Note: RCPT_AGE- age at opioid dispensation; Fluiddo- fluid disorder according to Elixhauser co-morbidity; Gender_M-male sex' NO5B-ANXIOLYTICS- prescribed ATC code benzodiazepine derivatives; Pharmacy_Risk_30- derived feature using proportion of opioid/benzodiazepine patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each pharmacy; N03A-ANTIEPILEPTICS- ATC code for anti-epileptics dispensed to patient; Doctor_Risk_30- derived feature using proportion of opioid/benzodiazepine patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each physician; Num_Opi_Fills_30- number of opioid dispensations in the previous 30 days prior to opioid dispensation; PriorHospRelevant- flag for history of any opioid related hospitalization in the previous 180 days prior to opioid dispensation; N07B-DRUGS USED IN ADDICTIVE DISORDERS- ATC code for drugs dispensed to patient for treating substance abuse disorders; A11D-VITAMIN B1, PLAIN AND IN COMBINATION WITH VITAMIN B6 AND B12- ATC code for patients dispensed Vitamins B1, B6, or B12; CODEIN: history of codeine use

**(C)**


**Note:** The "reference point" is called the "base value" at -3.902. Values in bold to the left of the base value indicate a lower predicted probability of the study outcome and values in bold to the right indicate a higher predicted probability of the study outcome. The top plot describes a patient at "low risk" for the study outcome. As can be seen from the feature values, this patient has a negative history for the specified features. The middle 2 plots describe a patient at "medium risk" while the bottom plot shows a patient at "high risk" for the study outcome.

**eFigure 4.** Calibration curve plotting observed vs. quantiles of estimated risk for the logistic regression (L1) classifier using the 2018 validation dataset. The majority of counts (dispensations) were predicted to be lower risk.



**eReferences**

1. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning, vol. 1. Springer series in statistics New York (2001)
2. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
3. Zhu, H. Zou, S. Rosset, T. Hastie, "Multi-class AdaBoost", 2009.
4. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2011 May 6;2(3):1-27.
5. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
6. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. InProceedings of the 27th international conference on machine learning (ICML-10) 2010 (pp. 807-814).
7. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).

BMJ Open

# Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada

| | |
|---|---|
| Journal: | *BMJ Open* |
| Manuscript ID | bmjopen-2020-043964.R2 |
| Article Type: | Original research |
| Date Submitted by the Author: | 12-Apr-2021 |
| Complete List of Authors: | Sharma, Vishal; University of Alberta, School of Public Health<br>Kulkarni, Vinaykumar; OKAKI Health Analytics<br>Eurich, Dean; University of Alberta, School of Public Health<br>Kumar, Luke; Alberta Machine Intelligence Institute<br>Samanani, Salim; Okaki Health Intelligence, |
| Primary Subject Heading: | Epidemiology |
| Secondary Subject Heading: | Health informatics, Public health |
| Keywords: | PUBLIC HEALTH, EPIDEMIOLOGY, Adverse events < THERAPEUTICS, Health & safety < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Clinical governance < HEALTH SERVICES ADMINISTRATION & MANAGEMENT |
| | |

*I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.*

*The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.*

*Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.*

**Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada**

**Author list (in order):**

Vishal Sharma (0000-0001-7907-1183), Vinaykumar Kulkarni, Dean T. Eurich (0000-0003-2197-0463), Luke Kumar, Salim Samanani (0000-0001-6751-4805)

**Address for each author:**

2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of Alberta, Edmonton, Alberta, Canada, T6G 2E1 Vishal Sharma BPharm PhD Candidate,

OKAKI Health Intelligence, Edmonton, Alberta, Canada, Vinaykumar Kulkarni MSc

2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of Alberta, Edmonton, Alberta, Canada, T6G 2E1 Dean Eurich professor

Alberta Machine Intelligence Institute, Edmonton, Alberta, Canada, T5J 3B1 Luke Kumar MSc

OKAKI Health Intelligence, Calgary, Alberta, Canada, Salim Samanani MD, Medical Director

**Corresponding Author:**

Dean Eurich, 2-040 Li Ka Shing Center for Health Research Innovation, University of Alberta, Edmonton, Alberta, Canada, T6G 2E1; Phone 780-492-6333; fax 780-492-7455; email: deurich@ualberta.ca

**Acknowledgement**

This study is based on data provided by The Alberta Strategy for Patient Orientated Research (AbSPORU) SUPPORT unit and Alberta Health. The interpretation and conclusions contained herein are those of the researchers and do not necessarily represent the views of the Government of Alberta or AbSPOR. Neither the Government of Alberta, ABSPOR nor Alberta Health expresses any opinion in relation to this study. This work was supported by Mitacs through the Mitacs Accelerate Program (VS and DTE).

**Contributors:** VS VK LK SS and DTE were involved in the conception and design of the study. VS VK LK SS and DTE analyzed the data. VS VK and LK drafted the article. VS VK LK DTE and SS revised the article. All authors gave final approval of the version to be published. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. DTE is the guarantor.

**Funding:** This study received no funding.



***Competing Interest:*** *All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; Salim Samanani has received grants from the College of Physicians & Surgeons of Alberta; no other relationships or activities that could appear to have influenced the submitted work.*

**Ethical approval:** This study was approved by the Health Research Ethics board at the University of Alberta (#Pro00083807_AME2).

**Data Sharing:** The data used in this study is not available for external analysis. However, administrative health data can be accessed from Alberta Health by following defined research protocols and confidentiality agreements.

**Transparency:** The lead author, VS, (the manuscript's guarantor, Dean Eurich) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as originally planned (and, if relevant, registered) have been explained.

**Word Count: 3357**

## Abstract

**Objective:** To develop machine-learning models employing administrative-health data that can estimate risk of adverse outcomes within 30-days of an opioid dispensation for use by health-departments or prescription monitoring programs.

**Design, Setting, and Participants:** This prognostic study was conducted in Alberta, Canada between 2017-2018. Participants included all patients over 18 years of age who received at least one opioid dispensation. Pregnant and cancer patients were excluded.

**Exposure:** Each opioid dispensation served as an exposure.

**Main Outcomes/Measures:** Opioid related adverse outcomes were identified from linked administrative health-data. Machine-learning algorithms were trained using 2017 data to predict risk of hospitalization, emergency department visit, and mortality within 30-days of an opioid dispensation. Two validation sets, using 2017 and 2018 data, were used to evaluate model performance. Model discrimination and calibration performance were assessed for all patients and those at higher risk. Machine-learning discrimination was compared to current opioid guidelines.

**Results:** Participants in the 2017 training set (n=275,150) and validation set (n=117,829) had similar baseline characteristics. In the 2017 validation set, c-statistics for the XGBoost, logistic regression, and neural network classifiers were 0.87, 0.87, and 0.80, respectively. In the 2018 validation set (n=393,023), the corresponding c-statistics were 0.88, 0.88, and 0.82. C-statistics from the Canadian guidelines ranged from 0.54-0.69 while the US guidelines ranged from 0.50-0.62. The top 5-percentile of predicted risk for the XGBoost and logistic regression classifiers captured 42% of all events and translated into post-test probabilities of 13.38% and 13.45%, respectively, up from the pre-test probability of 1.6%.

**Conclusion:** Machine-learning classifiers, especially incorporating hospitalization/physician claims data, have better predictive performance compared to guideline or prescription history only approaches when predicting 30-day risk of adverse outcomes. Prescription monitoring programs and health departments with access to administrative data can use machine-learning classifiers to effectively identify those at higher risk compared to current guideline-based approaches.

## Article Summary

### Strengths and Limitations:

- This study incorporated near complete capture of opioid dispensations from community pharmacies and used validated administrative health data.
- This study used commonly available algorithms to train machine-learning models using data which is available to government health departments in all provinces in Canada and other single payer jurisdictions; ML classifiers were evaluated with informative prognostic metrics not usually seen in other studies.
- Our predictive models used dispense events and not medication utilization, which is difficult to capture in administrative data.
- Our training dataset does not account for non-prescription opioids, opioids administered in hospitals, and other risks associated with non-prescription use.

## Introduction

Canada is among the countries with the highest rates of opioid prescribing in the world, making prescription opioid use a key driver of the current opioid crisis[1]; a major part of the policy response to the opioid crisis focuses on endorsing safe, appropriate opioid prescribing[2-4]. In order to minimize high risk opioid prescribing and to identify patients at high risk of opioid related adverse outcomes, numerous health regulatory bodies have released clinical practice recommendations for health providers regarding appropriate opioid prescribing[3,5,6].

Prescription monitoring programs (PMPs) have been implemented around the world, like Alberta's provincial Triplicate Prescription Program (TPP)[7] in Canada, and are mandated to monitor the utilization and appropriate use of opioids to reduce adverse outcomes. In most jurisdictions, both population-level monitoring metrics and clinical decision aids are used to identify patients at risk of hospitalization or death and are most often based on prescribing guidelines. However, a comprehensive infrastructure of administrative data containing patient level International Statistical Classification of Diseases and Related Health Problems (ICD)[8] codes and prescription drug histories exists in Alberta and other provinces in Canada which could be further integrated to predict opioid-related risk. Furthermore, current guidelines addressing high risk prescribing and utilization of opioids were derived from studies that used traditional statistical methods to identify population level risk factors for overdose rather than an individual's absolute risk[3,9,10]; these population estimates may not be generalizable to different populations[11]. Thus, a functional gap exists in many health jurisdictions where much of the available administrative health data is not being leveraged for opioid prescription monitoring.

Supervised machine learning (ML)[12,13] is an approach that uses computer algorithms to build predictive models in the clinical setting that can make use of the large amounts of available administrative data[14,15], all within a well-defined process[16]. Supervised ML trains on labelled data to develop prediction models that are specific to different populations and, in many cases, can provide better predictive performance than traditional, population-based statistical models[10,15,17]. We identified one study[10] that applied ML techniques to predict overdose risk in opioid patients pursuant to a prescription. In their validation sample, they found that the deep neural network (DNN) and gradient boosting machines (GBM) algorithms carried the best discrimination performance based on estimated c-statistics and that the ML approach out-performed the guideline approach in terms of risk prediction; neural networks have little interpretability and are not necessarily better at predicting outcomes when trained on structured data[18]. This study relied on c-statistics to evaluate their ML models and did not emphasize other performance metrics required to assess clinical utility that are recommended by medical reporting guidelines[11,13,19,20]. It also did not address the important issue of ML model interpretability[21]. Reporting informative prognostic metrics is needed to better understand the capabilities of ML classifiers if health departments and PMPs are to incorporate them into their decision-making processes.

The objective of our study was to further develop and validate ML algorithms (beyond just DNN) to predict the 30-day risk of hospitalization, emergency visit and mortality for a patient in Alberta, Canada at the time of an opioid dispensation using administrative data routinely available to health departments and PMPs and evaluate them using the above referenced reporting guidelines. We also analyzed feature importance to provide meaningful

interpretations of the ML models. Comparing discrimination performance (area under the receiver operating characteristics curves), we hypothesized that the ML process would perform better than the current guideline approach for predicting risk of adverse outcomes related to opioid prescribing.

## Methods

### Study Design and Participants

This prognostic study used a supervised ML scheme. All patients in Alberta, Canada who received a dispensation for an opioid, were 18 years of age and older between Jan 1, 2017 and Dec 31, 2018 were eligible. Patients were excluded from all analyses if they had any previous diagnosis of cancer, received palliative interventions or were pregnant during the study period (eTable 1 in Supplement) as use of opioids in these contexts is clinically different.

Government health departments and payers in many jurisdictions have systems to capture prescription histories and ICD diagnostic codes. As such, we linked various administrative health data sets available in Alberta, Canada using unique patient identifiers in order to establish a complete description of patient demographics, drug exposures and health outcomes. These databases include *1) Pharmaceutical Information Network (PIN):* PIN data includes all dispensing records from community pharmacies from all prescriber types occurring in the province outside of the hospital setting. PIN collects all drug dispensations irrespective of age or insurance status in Alberta; Anatomical Therapeutic Chemical classification (ATC) codes[22] were used to identify opioid dispensations (eSupplement), 2*) Population and Vital Statistics Data (VS, Alberta Services)*: sex, age, date of birth, death date, immigration and emigration

data, and underlying cause of death according to the World Health Organization algorithm using ICD codes[8], *3) Hospitalizations and Emergency Department Visits (National Ambulatory Care Reporting System [NACRS], Discharge Abstract Database [DAD]):* all services, length of stay, diagnosis (up to 25 ICD-10[8] based diagnoses). Data and coding accuracy are routinely validated both provincially and centrally via the Canadian Institute for Health Information, and 4*) Physician Visits/Claims (Alberta Health):* all claims from all settings (e.g., outpatient, office visits, emergency departments, inpatient) with associated date of service, ICD code, procedure and billing information.

This study followed the TRIPOD and STARD reporting guidelines[23-25] and received ethics approval from the University of Alberta ethics board (Pro00083807_AME1). All analyses were done using Python (v. 3.6.8,), SciKit Learn[26] (v. 0.23.2) SHAP[27] (v. 0.35), XGBoost (v. 0.90)[28], Pandas (v. 1.0.5)[29] and H20 Driverless AI (version 1.9).

**Measures and Outcome**

ML models were trained on a labelled dataset in which the observation/analysis unit was an opioid dispensation. Every opioid dispensation, not just the incident one, was used as a potential instance to predict the risk of our outcome. The primary outcome was a composite of a drug-related hospitalization, emergency department (ED) visit or mortality within 30 days of an opioid dispensation based on ICD-10 codes identified from DAD, NACRS and Vital Statistics (T40, F55, F10-19; eTable 2 in Supplement)[2,10,30].

We anticipated that our defined outcome would be a rare event, leading to a class imbalanced dataset[31]. To address this, we relied on specifying balanced class weightage for

supporting algorithms; other approaches were deemed not suitable (e.g., oversampling using randomly repeating minority class); under sampling (sub-sampling within the majority class) resulted in changes in outcome prevalence. Class weightage is a commonly used method[32] to address class imbalance along with over and under-sampling approaches. However, oversampling, which involves generating new opioid dispensations from the original data distribution and is prone to introducing bias, is difficult due to the categorical nature of the data and beyond the scope of this study. With under-sampling, which takes samples from the majority class (in this case, no 30-day event after dispensation), we would not be able to use all of the information provided by the data in instances with no outcome. Hence, we decided to use the class weightage method which does not alter the data distribution. Instead, the learning process is adjusted in a way that increases the importance of the positive class (instances that led to a 30-day event)[33].

**Predictor Candidates for ML Models**

Predictor variables in our ML models included those that were informed by the literature[3,4,10] and those directly obtained from the data sets. These included features based on demographics (age, sex, income using Forward Sortation index from postal codes[34]), co-morbidity history using ICD-based Elixhauser score categories[35], health care utilization (number of unique providers, number of hospital and emergency department visits), and drug utilization (level 3 ATC codes[22], oral morphine equivalents[36], concurrent use with benzodiazepines, number of opioid and benzodiazepine dispensations, number of unique opioid and benzodiazepine molecules). Depending on the potential predictor and data availability, we used data from 30 days to 5 years before the opioid dispensation to generate model features

(eFigure 1 in Supplement); 30 days was used to reflect the immediate nature of the risk and 5 years to fully capture co-morbidities. This approach aligns with how health providers would assess patients using the entire history of co-morbidities and then the more immediate factors in deciding on the need for a therapeutic as well as risk in patients. We performed experiments to identify the features and data sets that contributed most to predicting the outcomes with a view to minimizing the potential future data requirements for health departments and PMPs.

**Statistical Analyses and Machine-Learning Prediction Evaluation**

We randomly divided the patients in the 2017 portion of our study cohort into training (70%) and validation (30%) sets[13] by patients and opioid dispensations such that no patients in the training set were in the validation set. Baseline characteristics and event rates were compared in the training vs validation group, and between those who experienced the outcome and those who did not using chi-squared tests of independence. As well, we used all the 2018 data as another independent validation set.

We trained commonly used[13,37] ML algorithms (eAppendix in Supplement) and further tuned out-of-box models using 5-fold cross validation on the training data to address model overfitting[13,38]. As is common in ML validation studies[10,13], we reported model discrimination performance (i.e. how well a model differentiates those at higher risk from those at lower risk)[11] using area under the receiver operating characteristic curve (AUROC; c-statistic). We then stratified the two ML models with the highest c-statistics into percentile categories (deciles) according to absolute risk of our outcome, as was done in previous studies[10,39]. We also plotted AUROC[11] and precision-recall curves (PRCs)[40].

Because discrimination alone is insufficient to assess ML model prediction capability, we assessed a second necessary property, namely, calibration (i.e., how similar the predicted absolute risk is to the observed risk across different risk strata)[11,41]. Using the two ML models with the highest discrimination performance, we assessed calibration performance on the 2018 data by plotting observed (fraction of positives) vs predicted risk (mean predicted value). Using these same two ML classifiers, we analyzed the top 0.1, 1, 5, and 10 percentiles of predicted risk by the number of true and false positives, positive likelihood ratios (PLR)[20], positive predictive values (PPV), post-test probabilities, and number needed to screen. We also performed a simulation of daily data uploads for 2018 Quarter 1 to view the predictive capabilities if a ML risk predictor were to be deployed into a monitoring workflow.

For the XGBoost and logistic regression classifiers, we reported feature importance[37] and plotted PRCs that compared all dispenses to those within the top 10 percentiles of estimated risk. As well, for the XGBoost classifier, we described feature importance on model outcome using SHAP values[27,42] to add an additional layer of interpretability.

Finally, we compared ML risk prediction (the two ML models with highest discrimination performance) to current guideline approaches as others have[10], using the 2019 Centers for Medicare & Medicaid Services (CMS) opioid safety measures[43] and the 2017 Canadian Opioid Prescribing Guideline[3]. We also compared the discrimination performance of different logistic regression classifier models using various combinations of features derived from their respective databases: **1)** demographic and drug/health utilization features from PIN and 2**)** co-morbidity features derived from DAD, NACRS and Claims.

## Patient and Public Involvement

This research was done without patient involvement. Patients were not invited to comment on the study design and were not consulted to develop patient relevant outcomes or interpret the results. Patients were not invited to contribute to the writing or editing of this document for readability or accuracy. There are no plans to disseminate the results of the research to study participants.

## Results

### Patient Characteristics and Predictors

We identified 392,979 patients with at least one opioid dispensation in 2017 (Figure 1). This cohort was used to train (n= 275,150, 70%) and validate (n=117,829, 30%) ML models. In 2017 and 2018, 6,608 and 5,423 patients experienced the defined outcome, respectively. Baseline characteristics were different between those who experienced the outcome and those who did not (eTable 3 in Supplement) while characteristics were similar between the training and validation sets (eTable 4 in Supplement). There were 2,283,075 opioid dispensations in 2017 and 1,977,389 in 2018. Overall, in 2017, 2.03% (n= 45,757) of opioid dispensations were associated with the outcome; in 2018, the estimate was 1.6% (n= 31,392).

As described above, we categorized our candidate features into four groups (eTable 5 in Supplement). When using all of the databases, the total number of features was 283 and 34 when considering only co-morbidities.

### Machine-Learning Prediction Performance

Using the 2017 validation set, AUROCs for the XGBoost and logistic regression classifiers had the highest discrimination performance at 0.87, while the neural network classifier had lower performance at 0.80 (eTable 6 in Supplement).

Discrimination performance was similar for the 2018 validation set (n=393,023; eTable 6 in Supplement). XGBoost and logistic regression had the highest estimated AUROCs and area under PRCs while the neural network classifier was lower (Figure 2A, 2B). As expected, precision-recall curves indicate stronger predictive performance in opioid dispensations at higher predicted risk percentiles (Figure 2C, 2D).

In the 2018 validation set, although discrimination performance was similar (0.88), individual feature importance was different between the logistic regression and XGBoost classifiers, with logistic regression feature importance more reliant on co-morbidity data from DAD, NACRS and Claims while XGBoost relied more on drug utilization data from PIN (eFigure 2). With the XGBoost classifier, history of drug abuse, alcoholism, and prior hospitalization/emergency visit carried the highest importance for predicting the study outcome (eFigure 3A) where the presence of these features in a patient suggested a strong prediction towards having the defined outcome (eFigure 3B and 3C).

**Calibration**

When considering dispensations predicted to be in the highest percentiles of risk, the top 5-percentile captured 42% of all outcomes using the XGBoost and logistic regression classifiers (Table 1). Also, as the predicted risk percentiles get higher (top 10 percentile to top 0.1 percentile), so too do the corresponding PPVs with the top 0.1 percentile associated with a

PPV of 33% for the XGBoost classifier. As well, lower categories of risk percentiles were associated with lower outcomes (Figure 3, eFigure 4). When we simulated a monitoring workflow scenario with daily data uploads, a similar pattern was illustrated where the dispensations predicted to be higher risk had higher event rates (Figure 4).

After using the XGBoost and logistic regression classifiers to identify the dispensations in the highest predicted risk percentiles, the pre-test probability of the outcome (1.6%) was transformed into higher post-test probabilities, with higher probabilities in the riskier percentiles (Table 1). The number needed to screen also decreased as predicted risk increased (Table 1).

Comparing discrimination performance, ML risk prediction outperformed the current guideline approaches when using various combinations of guideline recommendations (Table 2). In many of the guideline scenarios, the estimated AUROCs were close to the 0.5 mark. When we estimated the discrimination performance of the logistic regression classifier based on database source, using all databases produced an AUROC of 0.88. Reducing the database source to only DAD, NACRS, Claims (co-morbidities only) resulted in an AUROC of 0.85, while PIN (prescription history) only was 0.78 (Table 3).

**Discussion**

This study showed that ML techniques using available administrative data (prescription histories and ICD codes) may provide enough discriminatory performance to predict adverse outcomes associated with opioid prescribing. Indeed, our ML analyses showed very high discrimination performance at 0.88. The linear model (logistic regression) and XGBoost carried

higher discrimination and calibration performance, while the neural network classifier did not perform as well. By identifying the predicted top 5-10 percentile of absolute risk pursuant to an opioid dispensation, we were able to capture approximately half of all outcomes using ML methods. All ML models we trained had higher discrimination performance using the validation sets compared to the clinical guideline approach.

Since the prevalence of our defined outcome is relatively low in the general population, PPVs would also be expectedly low. However, estimated PPVs increased when we considered higher risk dispensations, as is expected since PPV is related to event prevalence. This is important because different users of a risk predictor will require different predictive capabilities. Similarly, our estimates of positive likelihood ratios and associated post-test probabilities also increased in dispensations with higher predicted risk indicating the strong predictive capabilities of the XGBoost and logistic regression classifiers; likelihood ratios >10 generate conclusive changes from pre-test to post-test probabilities[20].

The current guideline approach to assess absolute opioid prescribing risk produced c-statistic estimates closer to 0.5 indicating that discrimination was not much better than chance alone. ML models with higher predictive performance can better support health departments and PMPs with monitoring mandates to identify and intervene on those at high risk and their associated prescribers. We also found that adding co-morbidity features from administrative databases increased prediction performance compared to prescription history alone, thus making the case for the use of this data by PMPs and health departments. However, if only prescription history is available, our trained XGBoost classifier still had strong discrimination performance.

We found only one study that used ML approaches to quantify the absolute risk of an event pursuant to an opioid dispensation[10]. Their methodology used rolling 3-month windows for estimating risk and ML model training while we used historic records to estimate 30-day risk. Differences in study population and feature selection may explain why their highest performing ML model was deep learning (neural network classifier) and ours was not. Nevertheless, we were able to replicate their predictive performance using our ML approach as we both showed that ML approaches have higher predictive capabilities than guideline approaches. Both of our studies used predicted percentile risk estimates to identify high risk dispensations and were able to do so with strong discrimination and calibration performance. Furthermore, we emphasized prognostic metrics which are more informative to assess the clinical utility of ML classifiers using pre- and post-test probabilities, something not done in other studies and recommended in medical guidelines[20]. This major aspect of our study, not done previously, is important because any ML classifier that does not increase prognostic information compared to baseline cannot be incorporated into decision making for the purpose of intervening on higher risk instead of lower risk patients. Indeed, another study we found describes how identifying cases in higher predicted risk percentiles using ML methods can be deployed in hospital settings for the purpose of targeted interventions[39] upon discharge, however the effect on outcomes is still to be determined.

The limitations of our study are similar to other ML studies[10] and need to be addressed when considering deployment of ML risk predictors. Our training dataset was not able to account for non-prescription opioid consumption and the risk associated with non-prescription use, both of which are substantial contributors to overall risk[2]. Regarding our analysis, we

assumed that all dispensations were independent events; future research in this area should focus on employing ML methods using correlated data. As with all ML projects, our models were trained using Alberta data and might not be generalizable to other populations, or to specific populations within Alberta. However, one of the benefits of the ML process is that models can be retrained or similar methods could be used to develop new models to accommodate different populations.

This study suggests that ML risk prediction can support PMPs, especially if readily available administrative health data is used. PMPs currently use population-based guidelines which we, and others, have shown cannot predict absolute individual risk. The ML process allows for flexibility in model training, validation and deployment to specific settings in which, for the case of PMPs, high risk patients can be identified and targeted for intervention either at the patient or provider level. For example, a ML classifier can be trained on accessible data to create an aggregated list of "high risk" patients at regular time intervals to identify points of intervention. Moreover, ML classifiers can be retrained over time as changes in populations and trends in prescribing occur and are therefore specific to the population unlike broadly based guidelines. Further research can assess whether implementation of a ML-based monitoring system by PMPs leads to improved clinical outcomes within their own jurisdictions and whether other available features or feature reduction can yield sufficiently valid results for their own intended purposes.

## References


1. Belzak L, Halverson J. Evidence synthesis - The opioid crisis in Canada: a national perspective. *Health Promotion and Chronic Disease Prevention in Canada.* 2018;38(6):224-233.
2. Gomes T, Khuu W, Martins D, et al. Contributions of prescribed and non-prescribed opioids to opioid related deaths: population based cohort study in Ontario, Canada. *BMJ.* 2018;362:k3207.
3. Busse JW, Craigie S, Juurlink DN, et al. Guideline for opioid therapy and chronic noncancer pain. *Canadian Medical Association Journal.* 2017;189(18):E659-E666.
4. Dowell D. CDC guideline for prescribing opioids for chronic pain. 2016.
5. ismp Canada. Essential Clinical Skills for Opioid Prescribers. 2017; https://www.ismp-canada.org/download/OpioidStewardship/Opioid-Prescribing-Skills.pdf. Accessed Nov 2018.
6. Centre for Effective Practice. Management of Chronic Non Cancer Pain. 2017; thewellhealth.ca/cncp.
7. College of Physicians and Surgeons of Alberta. TPP Alberta – OME and DDD Conversion Factors. 2020; http://www.cpsa.ca/tpp/. Accessed Jun 2020.
8. World health Organization. Classification of Diseases (ICD). 2019; https://www.who.int/classifications/icd/icdonlineversions/en/. Accessed Jun 2020.
9. Gomes T, Mamdani MM, Dhalla IA, Paterson JM, Juurlink DN. Opioid Dose and Drug-Related Mortality in Patients With Nonmalignant PainOpioid Dose and Drug-related Mortality. *JAMA Internal Medicine.* 2011;171(7):686-691.
10. Lo-Ciganic W-H, Huang JL, Zhang HH, et al. Evaluation of Machine-Learning Algorithms for Predicting Opioid Overdose Risk Among Medicare Beneficiaries With Opioid Prescriptions. *JAMA network open.* 2019;2(3):e190968-e190968.
11. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. *JAMA.* 2017;318(14):1377-1384.
12. Shah NH, Milstein A, Bagley P, Steven C. Making Machine Learning Models Clinically Useful. *JAMA.* 2019;322(14):1351-1352.
13. Liu Y, Chen P-HC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA.* 2019;322(18):1806-1816.
14. Bastanlar Y, Ozuysal M. Introduction to machine learning. *Methods in molecular biology (Clifton, NJ).* 2014;1107:105-128.
15. Thottakkara P, Ozrazgat-Baslanti T, Hupf BB, et al. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PloS one.* 2016;11(5):e0155705.
16. Alberta Machine Intelligence Institute. Machine Learning Process Lifecycle. In:2019.
17. Hsich E, Gorodeski EZ, Blackstone EH, Ishwaran H, Lauer MS. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circulation: Cardiovascular Quality and Outcomes.* 2011;4(1):39-45.
18. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Paper presented at: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining2015.
19. Yusuf M, Atal I, Li J, et al. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. *BMJ open.* 2020;10(3):e034568.
20. Jaeschke R, Guyatt GH, Sackett DL, et al. Users' Guides to the Medical Literature: III. How to Use an Article About a Diagnostic Test B. What Are the Results and Will They Help Me in Caring for My Patients? *JAMA.* 1994;271(9):703-707.
21. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA.* 2018;320(21):2199-2200.

22. World Health Organization. International language for drug utilization research, ATC/DDD. 2020; https://www.whocc.no/. Accessed Jun 2020, 2020.
23. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine.* 2015;162(1):W1-W73.
24. equator network. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. 2020; https://www.equator-network.org/reporting-guidelines/tripod-statement/. Accessed Feb 2020.
25. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open.* 2016;6(11):e012799.
26. Buitinck L, Louppe G, Blondel M, et al. API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:13090238.* 2013.
27. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Paper presented at: Advances in neural information processing systems2017.
28. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Paper presented at: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining2016.
29. The pandas development team. pandas-dev/pandas: Pandas. 2020; https://doi.org/10.5281/zenodo.3509134, Jan 2021.
30. Zhou H, Della PR, Roberts P, Goh L, Dhaliwal SS. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. *BMJ Open.* 2016;6(6):e011060.
31. Brownlee J. A Gentle Introduction to Imbalanced Classification. 2020; https://machinelearningmastery.com/what-is-imbalanced-classification/. Accessed Jan 2021.
32. King G, Zeng L. Logistic regression in rare events data. *Political analysis.* 2001;9(2):137-163.
33. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. *Journal of Big Data.* 2019;6(1):1-54.
34. Government of Canada. Forward Sortation Area—Definition. 2015; https://www.ic.gc.ca/eic/site/bsf-osb.nsf/eng/br03396.html. Accessed April 2020, 2020.
35. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical care.* 2005:1130-1139.
36. College of Physicians and Surgeons of Alberta. OME and DDD conversion factors. http://www.cpsa.ca/wp-content/uploads/2017/06/OME-and-DDD-Conversion-Factors.pdf.
37. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European heart journal.* 2017;38(23):1805-1814.
38. Rose S. Machine Learning for Prediction in Electronic Health Data. *JAMA Network Open.* 2018;1(4):e181404-e181404.
39. Morgan DJ, Bame B, Zimand P, et al. Assessment of Machine Learning vs Standard Prediction Rules for Predicting Hospital Readmissions. *JAMA Network Open.* 2019;2(3):e190348-e190348.
40. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE.* 2015;10(3):e0118432.
41. Shah ND, Steyerberg EW, Kent DM. Big Data and Predictive Analytics: Recalibrating Expectations. *JAMA.* 2018;320(1):27-28.
42. Molnar C. *Interpretable machine learning. A Guide for Making Black Box Models Explainable.* 2019.
43. Centers for Medicare & Medicaid Services (CMS). Announcement of calendar year (CY) 2019 Medicare Advantage capitation rates and Medicare Advantage and Part D payment policies and final call letter.

**Figure Legend**

Figure 1. Patient flow diagram of study participants used for training and validating ML models. NACRS: National Ambulatory Care Reporting System; DAD: Discharge Abstract Database; VS: Vital Statistics; PIN: Pharmaceutical Information Network; Claims: Physician Claims

Figure 2. Area under the receiver operating characteristic curve (AUROC) (A) and precision-recall curves (B) for all dispensations using logistic regression (L1), neural network, support vector machine (SVM), XGBoost and Naïve-Bayes; precision-recall curves for higher risk dispensations according to predicted risk percentile categories for logistic regression (C) and XGBoost (D) using the 2018 validation set.

Figure 3. Calibration curve plotting observed vs. quantiles (deciles) of estimated risk for the XGBoost classifier using the 2018 validation dataset. The majority of counts (dispensations) were predicted to be lower risk.

Figure 4. Simulation of a clinical workflow with daily uploads and events per 100 daily dispenses by risk percentiles using 2018 Quarter 1 (Q1) data for logistic regression (A) and XGBoost (B) classifiers.

Table 1. Highest percentiles of estimated risk and predictive performance using the XGBoost and logistic regression classifiers for the 2018 validation dataset (n=393,023). Total number of dispenses= 1,977,389; total number of outcomes= 31,392.

| Metric | Top 0.1%ile | | Top 1%ile | | Top 5%ile | | Top 10%ile | |
|---|---|---|---|---|---|---|---|---|
| | XGBoost | Logistic Regression | XGBoost | Logistic Regression | XGBoost | Logistic Regression | XGBoost | Logistic Regression |
| **Number of Dispenses** | 1,977 | 1,977 | 19,774 | 19,774 | 98,869 | 98,869 | 197,739 | 197,739 |
| **TP captured** | 655 | 472 | 4204 | 4100 | 13224 | 13293 | 18404 | 18409 |
| **Percent of TP** | 2.09 | 1.50 | 13.39 | 13.06 | 42.13 | 42.35 | 58.63 | 58.64 |
| **FP captured** | 1322 | 1505 | 15570 | 15674 | 85645 | 85576 | 179335 | 179330 |
| **PPV** | 33.13 | 23.87 | 21.26 | 20.73 | 13.38 | 13.45 | 9.31 | 9.31 |
| **PLR** | 30.71 | 19.44 | 16.74 | 16.22 | 9.57 | 9.63 | 6.36 | 6.36 |
| **Post-test Probability*** | 33.13 | 23.87 | 21.26 | 20.73 | 13.38 | 13.45 | 9.31 | 9.31 |
| **NNS** | 3.17 | 4.49 | 5.08 | 5.22 | 8.48 | 8.43 | 12.95 | 12.95 |

*Pre-test probability estimated at 1.6% using prevalence.

TP: true positives; FP: false positives; PPV: positive predictive value; PLR: positive likelihood ratio; NNS: number needed to screen

Note: Logistic regression used L1 (lasso) parameter regularization

Table 2. Discrimination performance of guideline approach using the 2018 validation set. Guideline approaches were adapted from the 2017 Canadian Opioid Prescribing Guideline and 2019 Centers for Medicare & Medicaid Services (CMS) opioid safety measures and compared to logistic regression and XGBoost classifiers (each with an estimated area under the receiver operating characteristic curve of 0.88).

| **Canadian Guidelines *** | **AUROC** | **Sensitivity** | **Specificity** |
|---|---|---|---|
| History of mental disorder only | 0.620 | 0.90 | 0.34 |
| Substance abuse only | 0.686 | 0.99 | 0.37 |
| OME/day >90 only | 0.539 | 0.22 | 0.85 |
| (Mental disorder and substance abuse) OR OME/day >90 | 0.690 | 0.91 | 0.47 |
| Mental disorder and substance abuse AND OME/day >90 | 0.560 | 0.20 | 0.91 |
| Mental disorder OR substance abuse OR OME/day >90 | 0.589 | 0.99 | 0.18 |
| **CMS Guidelines**** | | | |
| High opioid dose (>120 OME/day for 90+days) | 0.507 | 0.081 | 0.933 |
| Concurrency (Opioid & BZRA for 30+ days) | 0.575 | 0.423 | 0.727 |
| Multiple doctors (>4) | 0.591 | 0.294 | 0.888 |
| Multiple pharmacies (>4) | 0.537 | 0.120 | 0.959 |
| All conditions | 0.50 | 0.001 | 0.999 |
| Any condition | 0.622 | 0.62 | 0.625 |

OME: daily oral morphine equivalents; BZRA: benzodiazepine receptor agonist. Elixhauser scoring ICD codes were used to identify mental disorders and substance abuse.

*The Canadian guidelines do not specify timelines. >90 OME was determined by taking the average daily OME over the 30 days prior to dispensation

**The CMS guidelines specify a timeline of 90 or more days at >120 OME and concurrent use of opioids and benzodiazepines for 30 days or more

Table 3. Discrimination performance based on database source using area under the receiver operating characteristic curve (AUROC) for the logistic regression classifier on the 2018 validation set.

| **Database source** | **Predictor Variables formed from database** | **AUROC** | **Number of features** |
|---|---|---|---|
| PIN only | Drug utilization + Prescription history | 0.78 | 248* |
| DAD, NACRS, Claims | Co-morbidities | 0.85 | 34 |
| PIN, DAD NACRS, Claims (all databases used in study) | Demographic + Drug Utilization + Healthcare Utilization + Co-morbidities | 0.88 | 283 |

Note: drug utilization includes features describing oral morphine equivalents[36], concurrent use with benzodiazepines, number of opioid and benzodiazepine dispensations, number of unique opioid and benzodiazepine molecules; health care utilization includes features describing number of unique health providers visited, number of hospital/emergency department visits; logistic regression used L1 (lasso) parameter regularization

Figure 1. Patient flow diagram of study participants used for training and validating ML models. NACRS: National Ambulatory Care Reporting System; DAD: Discharge Abstract Database; VS: Vital Statistics; PIN: Pharmaceutical Information Network; Claims: Physician Claims

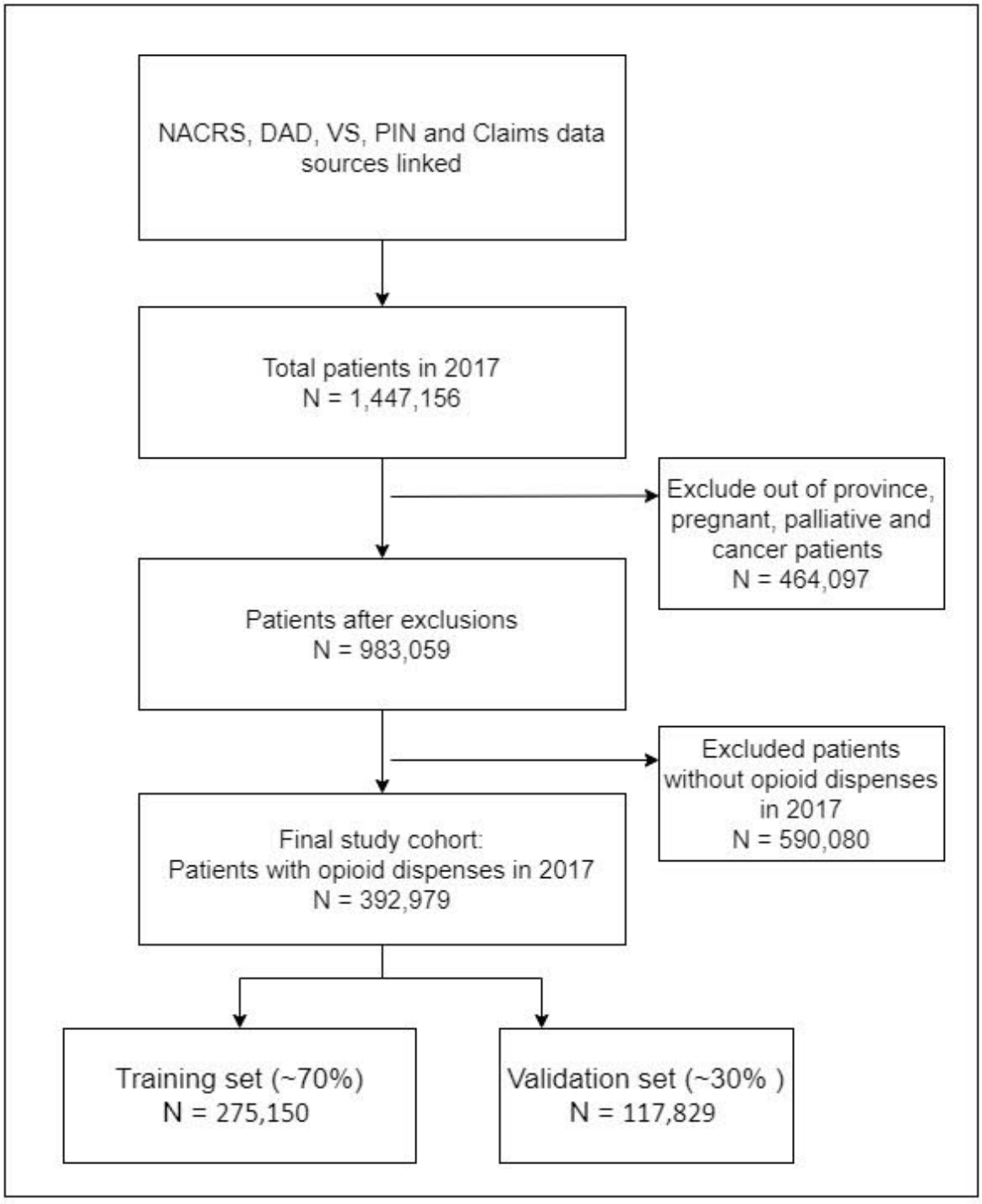
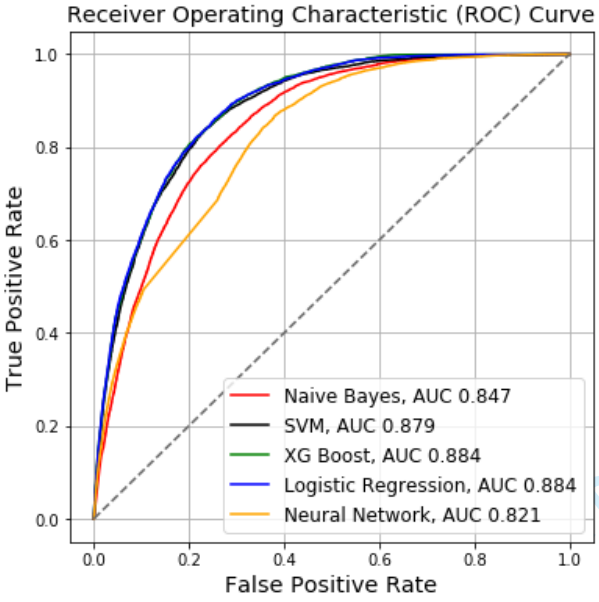
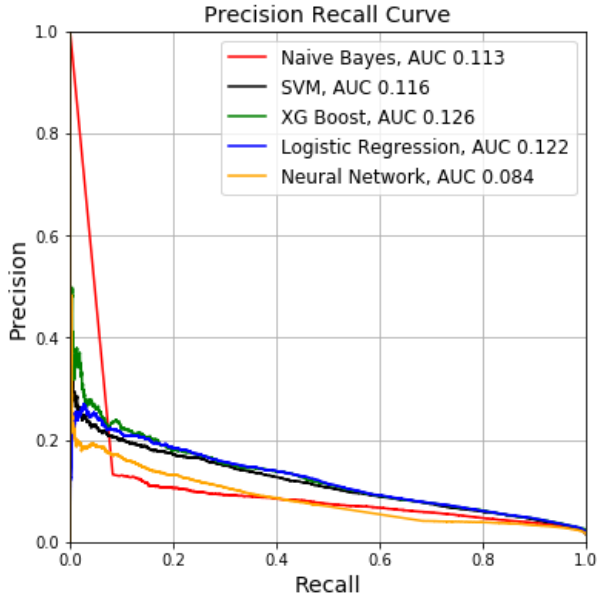


Figure 2. Area under the receiver operating characteristic curves (A) and precision-recall curves (B) for all dispensations using logistic regression (L1), neural network, support vector machine (SVM), XGBoost and Naïve-Bayes; precision-recall curves for higher risk dispensations according to predicted risk percentile categories for logistic regression (C) and XGBoost (D) using the 2018 validation set.

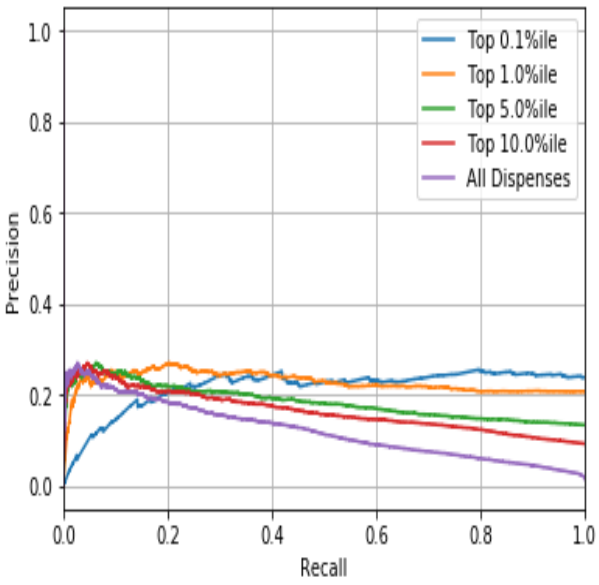
(A)



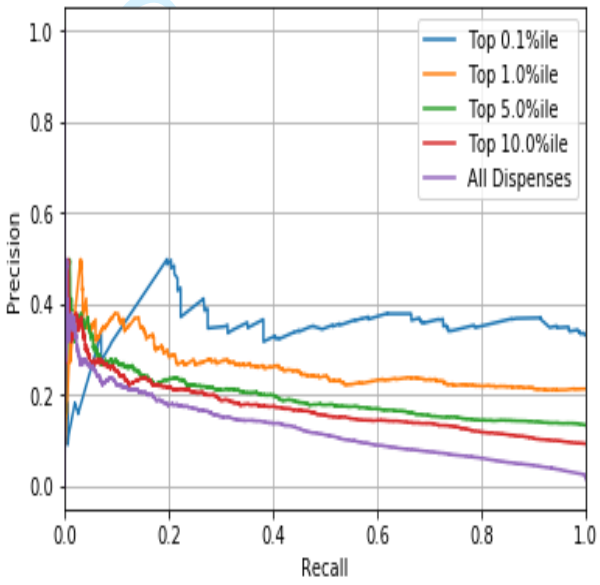

(B)




(C) Logistic Regression




(D) XGBoost




AUC: area under the curve

Figure 3. Calibration curve plotting observed vs. quantiles (deciles) of estimated risk for the XGBoost classifier using the 2018 validation dataset. The majority of counts (dispensations) were predicted to be lower risk.

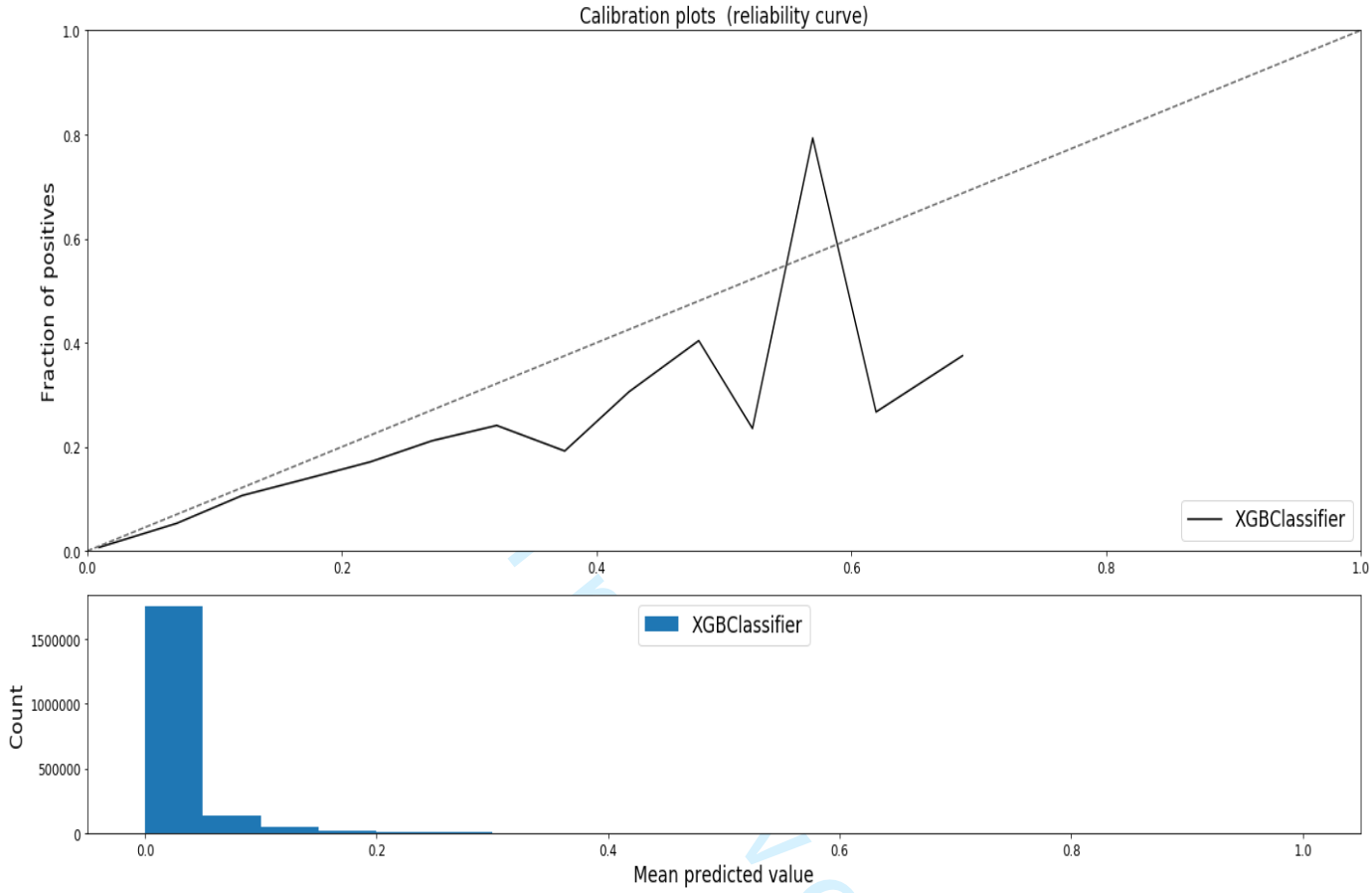
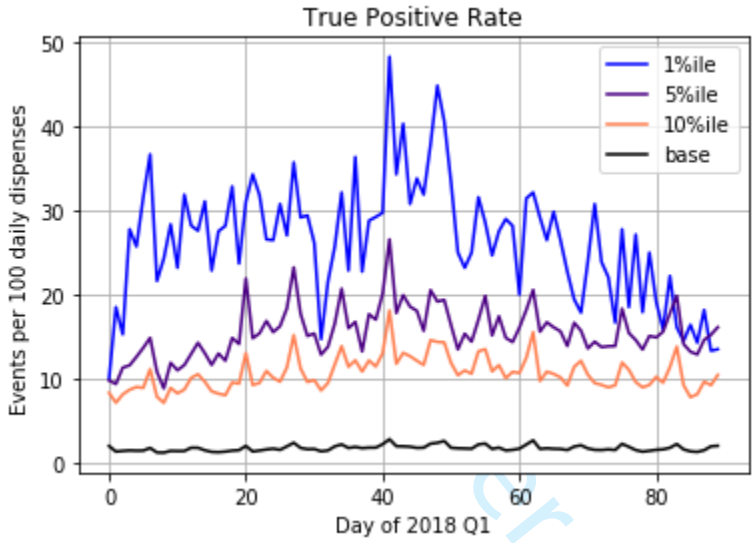
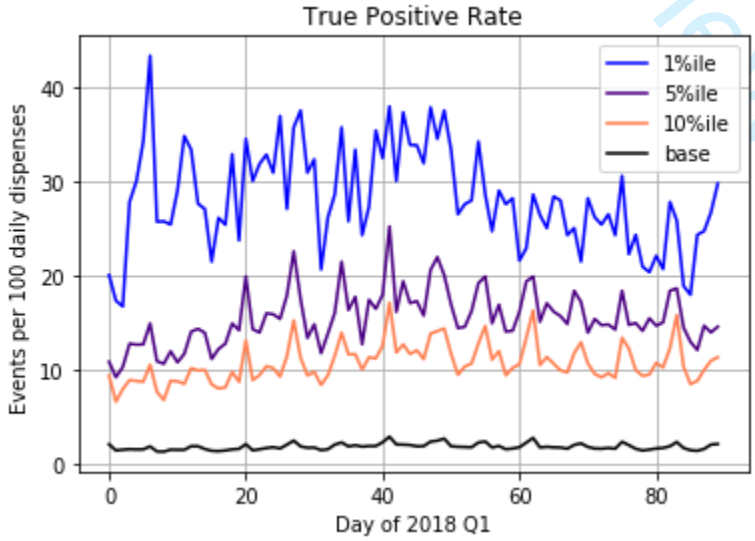


Figure 4. Simulation of a clinical workflow with daily uploads and events per 100 daily dispenses by risk percentiles using 2018 Quarter 1 (Q1) data for logistic regression (A) and XGBoost (B) classifiers.

(A) Logistic Regression (L1)




(B) XGBoost



**Supplementary Content**

**eAppendix.** Machine learning algorithms

**eTable 1.** Diagnostic codes used to exclude patients who had cancer, were pregnant, or were under palliative care.

**eTable 2.** Diagnostic codes used to identify the defined study outcome from emergency visit, hospitalization and death data.

**eTable 3.** Baseline characteristics of study patients (n=392,979). Co-morbidities were determined using Elixhauser criteria. All p-values in the $chi^2$ test of independence were <0.001 unless otherwise indicated.

**eTable 4.** Characteristics of study participants between training and validation groups using 2017 data.

**eTable 5.** Candidate predictors used to train ML algorithms.

**eTable 6.** Discrimination performance using area under the receiver operating characteristic curve (AUROC) of various ML algorithms. Training and validation were done using 2017 data (n=393,979); another independent validation was performed using 2018 data (n=393,023).

**eFigure1.** Schematic of study design and feature generation

**eFigure2**. Feature importance from logistic regression and tree-based (XGBoost) classifiers using the 2018 validation set.

**eFigure3**. Shapley values and feature impact in the XGBoost classifier using the 2018 validation set to describe "associations" between features and the outcome.

**eFigure 4.** Calibration curve plotting observed vs. quantiles of estimated risk for the logistic regression (L1) classifier using the 2018 validation dataset. The majority of counts (dispensations) were predicted to be lower risk.

**eReferences.**

# eAppendix. Machine Learning Algorithms

## Introduction

While there are always updates and new methods coming up in the fields of machine learning, in this study, we have focused on some of the most reliable and proven approaches for predictive modelling which are explainable and popularly used in previous studies of similar nature.

## Logistic Regression

Regression analysis models the relationship between a dependent variable and a set of independent variables [1]. Typically, this includes understanding how the value of the dependent variable changes with the changes in the values of independent variables. Logistic regression [1] uses the logistic function to model a binary dependent variable, where, based on the values of the independent variables the model can approximate one of the two classes, the instance belongs to. This basic binary model can be extended to deal with multiple classes (e.g. One-vs-all classifiers). However, logistic regression is only capable of modeling a linear relationship of independent variables to the dependent variable, hence limited to problems with linear decision boundaries. We used the sci-kit learn library in our experiments[6] and found L1 regularization to be more effective.

## Ridge Classifier

We used the ridge classifier implemented in the Scikit learn library[5]. It implements a classifier using ridge regression which uses an L2 regularization on the least square objective function. The library converts the labels into -1 and 1 and fits a linear regression on the converted labels with the regularization.

## Random Forest

Random forest is a tree ensemble learning algorithm that has wide applicability in many domains[1]. Random forest is a nonlinear learning algorithm, which arrives at nonlinear decision boundaries by independently combining multiple decision trees. Each individual decision tree in the forest can be grown independently of each other on a subset of the training data. Random forests are mainly sensitive to the number of trees, the depth of a tree and the number of covariates randomly chosen to split at each node[1]. These hyper-parameters can be tuned to find the best configuration of every dataset. Random Forests, in general, are less prone to overfit since they always grow individual trees on a subset of the training data[1]. At prediction time, the decision of each tree is aggregated to compute the final prediction.

## Neural Networks (NN)

Neural networks are another collection of non-linear learning algorithms with high representation power. They are known to be able to find mappings from an input to an output from a larger non-linear function space [2]. This ability to represent a larger space of nonlinear

functions has shown to be very effective recently in many application domains such as natural language processing, computer vision, genomics, computer games and health[2]. Neural networks come in many flavors learning nonlinear mapping of different types of data such as Convolutional NNs being most effective with images and Recurrent NNs for time series and language data. Identifying the most effective neural network structure is one of the difficult and the most time-consuming aspect of applying neural networks to new application domains and data. Generally, neural networks try to exploit the relationships in the raw unstructured data (eg: image and text) presented to the network but with more structured data such as health records and ICD codes learning relationships is much complex. Our neural network models are mainly based on densely connected hidden layers with ReLu[6] activation function. We used the cross-entropy loss for the binary classification Adam optimizer. We used a simple feed forward network using Sklearn MLP classifier with hyperparameter tuning for the NN.

**Boosted Learning Algorithms**

Boosting is a process to ensemble multiple base learning algorithms to arrive at better overall performance than any individual base learner[1]. In contrast to independently building multiple models from the subsets of the data, boosting re-weights the training data every time a model is learned for future models. This weighting happens to give more preference to currently misclassified data points in the next round compared to the correctly classified data points. Therefore future learners try to do better on the misclassified data points leading to a collection base learners having a better-combined prediction. This process is sequential so each base learner is dependent on the output of the previously trained model (it is worthy to note XGBoost provides a parallel tree boosting alternative). In our work, we have experimented with several boosting meta-learning algorithms such as XGBoost[7], AdaBoost[5] and GBM[5]. XGBoost uses a variant of trees as the base learner whereas AdaBoost (from Sci-kit learn) can use many ML algorithms as base learners. GBM uses logistic regression by default as the base learner. We used all 3 types of boosting with tuned hyperparameters for comparison.

**Naive Bayes**

Naive Bayes is based on the Bayes theorem with a strong independence assumption between the covariates[1]. This assumption helps in building a simple probabilistic model for learning and inference. Naive Bayes coefficients scale linearly with the number of covariates making this a suitable model for high-dimensional data. We used Naive Bayes as a simple baseline learning algorithm for comparison.

**Support Vector Machines (SVM)**

SVMs[4] are maximum margin classifiers optimizing for learning a hyperplane having the maximum distance away from each of the class data points[1]. SVM is a linear classifier but with the kernel trick to map the inputs to the higher dimensional space, it can learn nonlinear decision boundaries in the input space. SVMs are very effective binary classifiers with the kernel trick[1]. With larger datasets, SVMs tend to become more computationally intensive.

eTable 1. Diagnostic codes used to exclude patients who had cancer, were pregnant, or were under palliative care.

| **Condition** | **ICD 9** | **ICD 10** |
|---|---|---|
| **Cancer** | 140.x - 239.x | C00.x - C99.x, D00.x - D49.x |
| **Pregnancy** | 630.x - 679.x | O00.x - O99.x |
| **Palliative** | V66 | Z51.0, Z51.1, Z51.5 |

eTable 2. Diagnostic codes used to identify the defined study outcome from emergency visit, hospitalization and death data.

| **ICD 10** | **Condition** |
|---|---|
| **T40.x** | Poisoning by, adverse effect of and underdosing of narcotics and psychodysleptics |
| **F55.x** | Abuse of non-psychoactive substances |
| **F11.x - F19.x** | Mental and behavioral disorders due to psychoactive substance use |

eTable 3. Baseline characteristics of study patients (n=392,979). Co-morbidities were determined using Elixhauser criteria. All p-values in the chi$^2$ test of independence were <0.001 unless otherwise indicated.

| Characteristic | Number without Event n=386,371 | Percent | Number with Event n=6,608 | Percent |
|---|---|---|---|---|
| **Age:** | | | | |
| Mean (SD) | 48.1 (16.4) | -- | 41.2 (12.4) | -- |
| 18-45 | 162057 | 41.9 | 3466 | 52.4 |
| 45-65 | 154632 | 40.0 | 2656 | 40.2 |
| >65* | 69682 | 18.0 | 486 | 7.4 |
| **Male** | 197491 | 50.3 | 3922 | 59.4 |
| **Female** | 194794 | 49.7 | 2686 | 40.6 |
| **Alcohol Disorder** | 66320 | 16.9 | 5220 | 79.0 |
| **Arrhythmia** | 90621 | 23.1 | 1959 | 29.6 |
| **Blood Loss Anemia** | 1164 | 0.3 | 82 | 1.2 |
| **Congestive Heart Failure** | 18954 | 4.8 | 565 | 8.6 |
| **Coagulopathy** | 8053 | 2.1 | 356 | 5.4 |
| **Deficiency Anemia** | 34188 | 8.7 | 971 | 14.7 |
| **Depression** | 159140 | 40.6 | 5518 | 83.5 |
| **Diabetes**** | 64132 | 16.3 | 1408 | 21.3 |
| **Substance Abuse Disorder** | 74678 | 19.0 | 5485 | 83.0 |
| **Fluid Disorder** | 42690 | 10.9 | 3012 | 45.6 |
| **Hypertension**** | 140171 | 35.7 | 2624 | 39.7 |
| **Hypothyroidism** | 45519 | 11.6 | 601 | 9.1 |
| **Injury^** | 195688 | 49.9 | 5541 | 83.9 |
| **Liver Disorder** | 21656 | 5.5 | 1588 | 24.0 |
| **Neurologic Disorder** | 230490 | 58.8 | 5387 | 81.5 |
| **Obesity** | 63393 | 16.2 | 970 | 14.7 |
| **Poisoning^** | 17434 | 4.4 | 2775 | 42.0 |
| **Psychoses** | 35870 | 9.1 | 3162 | 47.9 |
| **Renal Disorder** | 16166 | 4.1 | 499 | 7.6 |
| **Rheumatoid Conditions** | 111458 | 28.4 | 3157 | 47.8 |
| **HIV Infection** | 1098 | 0.3 | 141 | 2.1 |
| **Paralysis** | 3874 | 1.0 | 187 | 2.8 |
| **Peptic Ulcer Disease** | 11728 | 3.0 | 509 | 7.7 |
| **Pulmonary Circulation Disorder** | 9611 | 2.4 | 430 | 6.5 |
| **Chronic Pulmonary Disease** | 102990 | 26.3 | 2913 | 44.1 |
| **Peripheral Vascular Disease** | 14467 | 3.7 | 389 | 5.9 |
| **Valvular Disease** | 7308 | 1.9 | 226 | 3.4 |
| **Weight Loss** | 16207 | 4.1 | 747 | 11.3 |

*p-value for age >65 is an estimated 0.037

^ Injury: ICD10: S00-T98; Poisoning: ICD10: T36-T50

** Complicated, uncomplicated diabetes and hypertension were collapsed into one category each

eTable 4. Characteristics of study participants between training and validation groups using 2017 data.

| Characteristic | Number in training group N=275,150~ | Percent | Number in validation group N=117,829~ | Percent |
|---|---|---|---|---|
| **Age:** | | | | |
| **Mean (SD)** | 48.3 (16) | -- | 48.2 (16) | -- |
| **18-45** | 114356 | 41.5 | 49909 | 42.3 |
| **45-65** | 111859 | 40.7 | 47132 | 40.0 |
| **>65** | 48935 | 17.8 | 20788 | 17.6 |
| **Male** | 138603 | 48.5 | 59339 | 48.4 |
| **Female** | 136545 | 47.8 | 58490 | 47.7 |
| **Alcohol Disorder** | 46792 | 16.4 | 20199 | 16.5 |
| **Arrhythmia** | 63637 | 22.3 | 27201 | 22.2 |
| **Blood Loss Anemia** | 839 | 0.3 | 336 | 0.3 |
| **Congestive Heart Failure** | 13320 | 4.7 | 5694 | 4.6 |
| **Coagulopathy** | 5697 | 2.0 | 2393 | 2.0 |
| **Deficiency Anemia** | 24096 | 8.4 | 10179 | 8.3 |
| **Depression** | 112080 | 39.2 | 47628 | 38.9 |
| **Diabetes**** | 45131 | 15.8 | 19144 | 15.6 |
| **Substance Abuse Disorder** | 52609 | 18.4 | 22713 | 18.5 |
| **Fluid Disorder** | 30272 | 10.6 | 12780 | 10.4 |
| **Hypertension**** | 98546 | 34.5 | 41840 | 34.1 |
| **Hypothyroidism** | 31908 | 11.2 | 13666 | 11.2 |
| **Injury*** | 137423 | 48.1 | 58865 | 48.0 |
| **Liver Disorder** | 15252 | 5.3 | 6567 | 5.4 |
| **Neurologic Disorder** | 161706 | 56.5 | 69341 | 56.6 |
| **Obesity** | 44607 | 15.6 | 18882 | 15.4 |
| **Poisoning*** | 12503 | 4.4 | 5293 | 4.3 |
| **Psychoses** | 25422 | 8.9 | 10860 | 8.9 |
| **Renal Disorder** | 11403 | 4.0 | 4817 | 3.9 |
| **Rheumatoid Conditions** | 78268 | 27.4 | 33420 | 27.3 |
| **HIV Infection** | 774 | 0.3 | 336 | 0.3 |
| **Paralysis** | 2717 | 1.0 | 1176 | 1.0 |
| **Peptic Ulcer Disease** | 8239 | 2.9 | 3533 | 2.9 |
| **Pulmonary Circulation Disorder** | 6771 | 2.4 | 2877 | 2.3 |
| **Chronic Pulmonary Disease** | 72265 | 25.3 | 30949 | 25.3 |

| | | | | |
|---|---|---|---|---|
| **Peripheral Vascular Disease** | 10228 | 3.6 | 4278 | 3.5 |
| **Valvular Disease** | 5111 | 1.8 | 2215 | 1.8 |
| **Weight Loss** | 11477 | 4.0 | 4790 | 3.9 |

Note: p-values for chi$^2$ test of independence were all >0.06 when comparing training and validation sets.
*Injury: ICD10: S00-T98; Poisoning: ICD10: T36-T50
** Complicated, uncomplicated diabetes and hypertension were collapsed into one category each

eTable 5. Anatomical Therapeutic Chemical classification of opioid molecules used for this study and candidate predictors used to train ML algorithms.

| **Category (data source)** | **Description** |
|---|---|
| ATC codes used to identify opioids from PIN data | N01AH01, N01AH03, N01AH06, N07BC01, N07BC02, N07BC51, R05DA03, R05DA04, R05DA09, R05DA20, N02A |
| Demographic information (PIN) | age, sex, postal codes, mean income |
| Drug utilization history (PIN) | drug dispenses in past 30 days using on ATC codes, oral morphine equivalents, concurrent use with benzodiazepines defined as at least 7 days of cumulative concurrent use in the 30 days prior to dispensation, number of dispensations and unique molecules of opioids and benzodiazepines |
| Health care utilization (PIN DAD) | flags for previous hospitalizations and emergency department visits, number of unique providers |
| ICD based co-morbidities (DAD, NACRS, Claims) | Elixhauser condition flags based on the past 5 years of claims, hospitalizations, and emergency visits. |

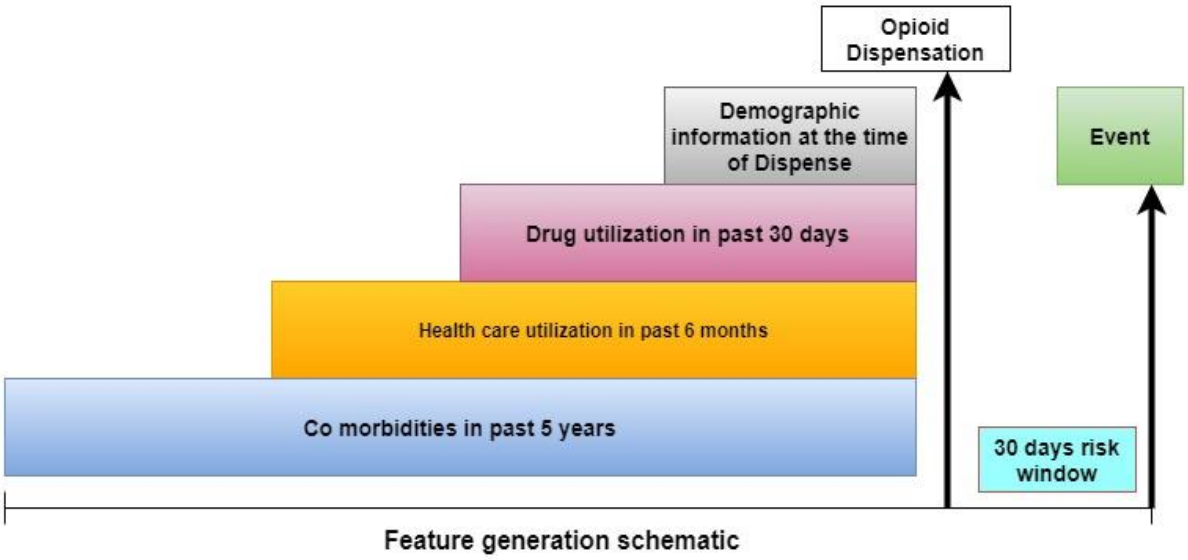
Note: ATC- Anatomical Therapeutic Chemical classification (https://www.whocc.no/atc_ddd_index); PIN- Pharmaceutical Information Network; ICD- International Statistical Classification of Diseases and Related Health Problems, World Health Organization; total number of features 283

eTable 6. Discrimination performance using area under the receiver operating characteristic curve (AUROC) of various ML algorithms using all features (demographics, health utilization, prescription history, co-morbidities). Training and validation were done using 2017 data (n=393,979); another independent validation was performed using 2018 data (n=393,023).

| Algorithm | Train | Validation 2017 | Validation 2018 |
|---|---|---|---|
| XGBoost Classifier | 0.897 | 0.870 | 0.884 |
| Logistic Regression | 0.887 | 0.869 | 0.884 |
| Gradient Boosting Classifier | 0.898 | 0.868 | 0.883 |
| AdaBoost Classifier | 0.884 | 0.868 | 0.882 |
| Random Forest Classifier | 0.909 | 0.863 | 0.881 |
| Ridge Classifier | 0.895 | 0.863 | 0.879 |
| SVM | 0.896 | 0.860 | 0.878 |
| Gaussian Naive Bayes | 0.846 | 0.826 | 0.847 |
| Decision Tree Classifier | 0.919 | 0.791 | 0.822 |
| Neural Networks | 0.827 | 0.804 | 0.821 |

Note: Logistic regression used L1 (lasso) parameter regularization

**eFigure 1. Schematic of study design and feature generation**



**eFigure2. Feature importance from logistic regression and tree-based XGBoost classifiers using the 2018 validation set. The logistic regression classifier relied more on co-morbidity data from DAD, NACRS, and Claims databases; XGBoost classifier relied more on data from the PIN database. AUROCs for both classifiers were similar at 0.88.**

| Logistic Regression | | XGBoost | |
|---|---|---|---|
| history of drug abuse | 1.00 | age at dispensation | 1.00 |
| age at dispensation | 0.65 | number of prescriptions dispensed in previous 30 days | 1.00 |
| history of prior hospitalization/ED visit | 0.62 | number of opioid dispensations in previous 30 days | 0.86 |
| history of alcohol use disorder | 0.62 | number of BZD dispensations in previous 30 days | 0.46 |
| history of fluid and electrolyte disorder | 0.32 | Doctor risk score* | 0.45 |
| history of poisoning | 0.31 | total OME consumed in previous 30 days | 0.43 |
| history of psychoses | 0.31 | history of poisoning | 0.37 |
| number of unique BZD dispensed in previous 30 days | 0.26 | pharmacy risk score** | 0.35 |
| history of depression | 0.19 | number of unique providers that prescribed an opioid or BZD | 0.34 |
| concurrent use of opioid and BZD in previous 30 days | 0.19 | income | 0.34 |
| history of injury | 0.17 | history of prior hospitalization/ED visit | 0.26 |

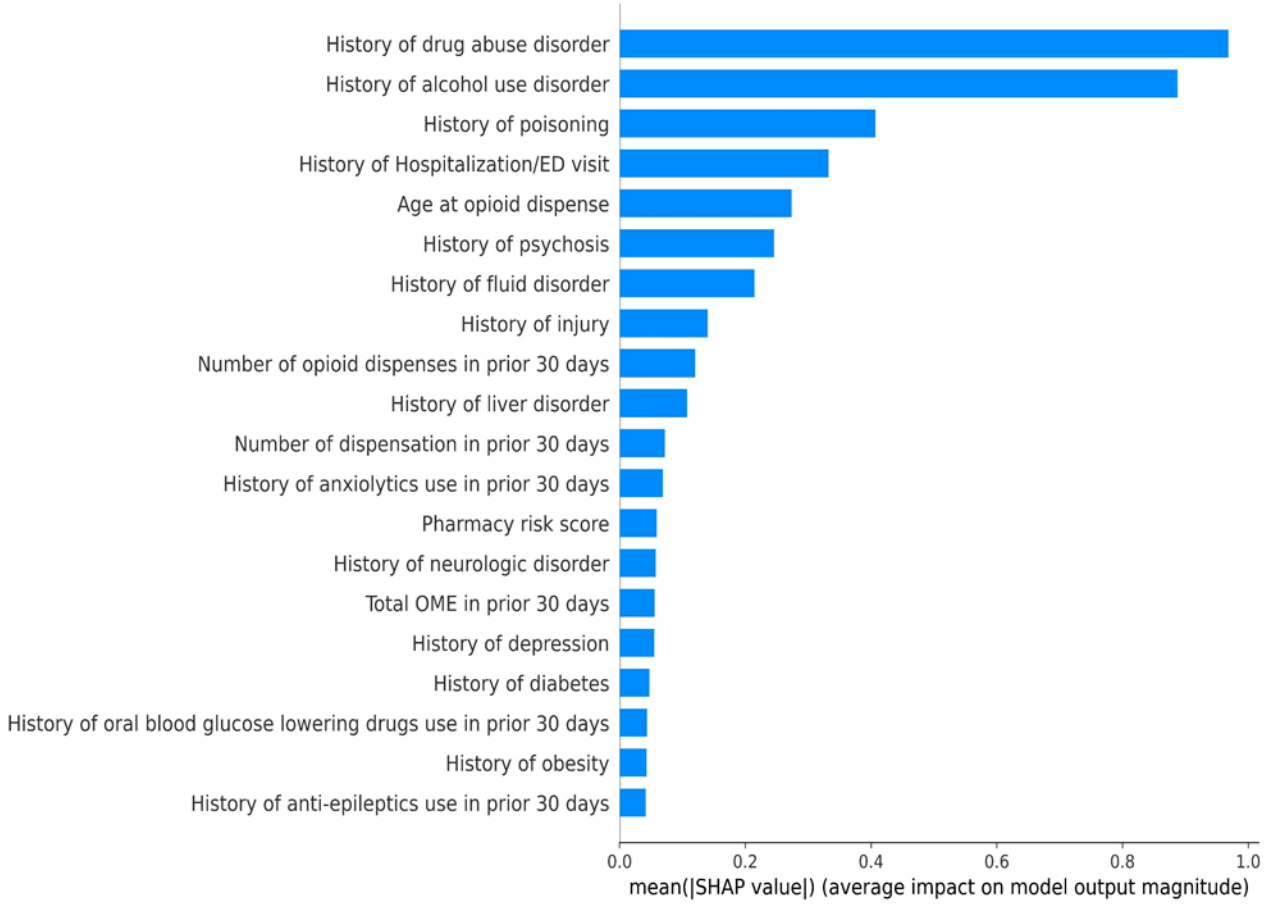
Note: Logistic regression used L1 (lasso) parameter regularization; BZD- benzodiazepine; OME- oral morphine equivalents; ED: emergency department

*derived feature using proportion of opioid/benzodiazepine patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each physician;

**derived feature using proportion of opioid/benzodiazepine patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each pharmacy

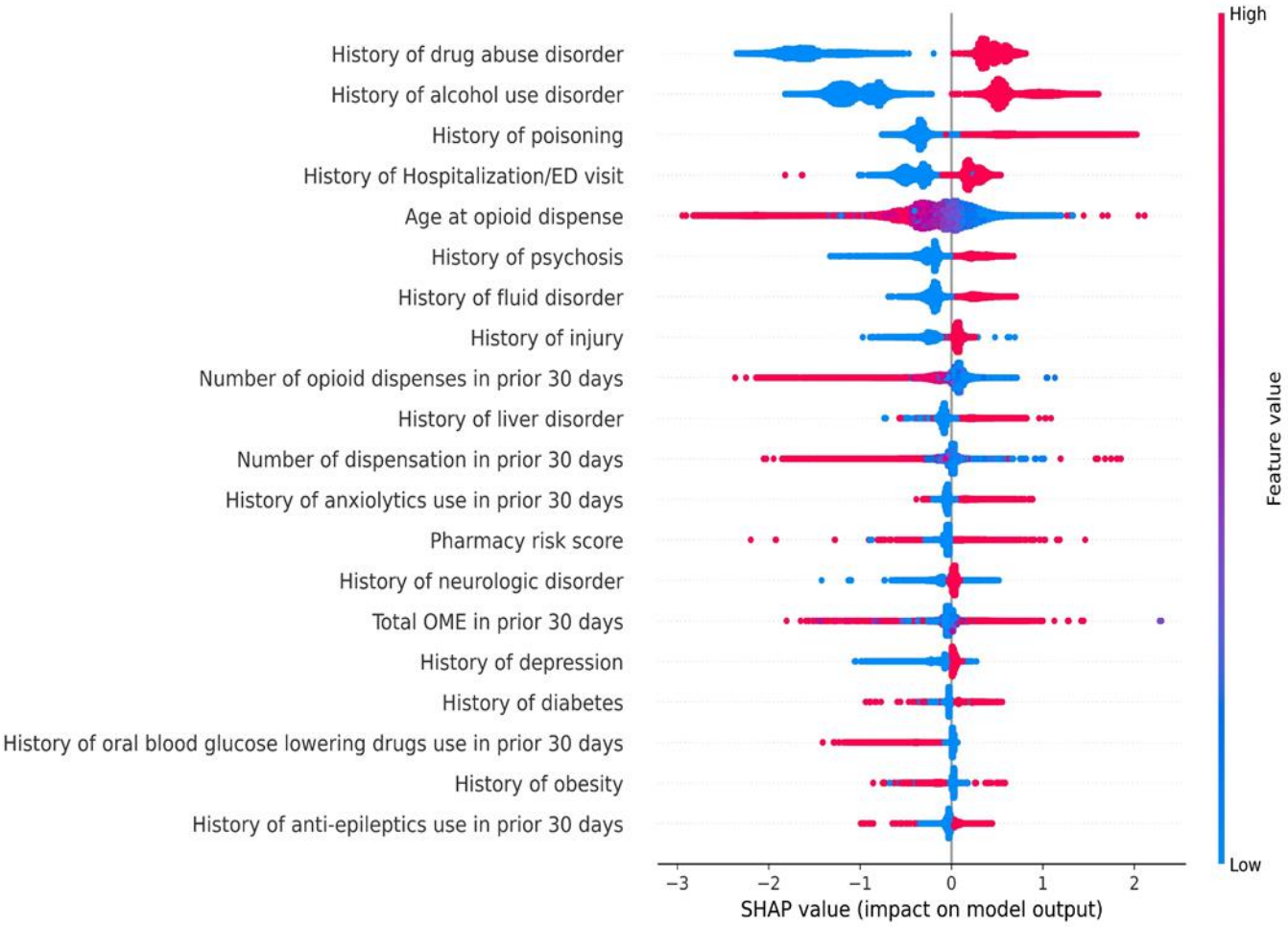
**eFigure 3. SHAP values and feature impact of the XGBoost classifier using the 2018 validation set to describe "associations" between features and the outcome. Features with the most impact on the model with drug abuse ranked highest (A); tornado plot illustrating feature impact (B); explaining the prediction of study outcome based on predictor values for 4 patients using SHAP values(C).**

**(A)**



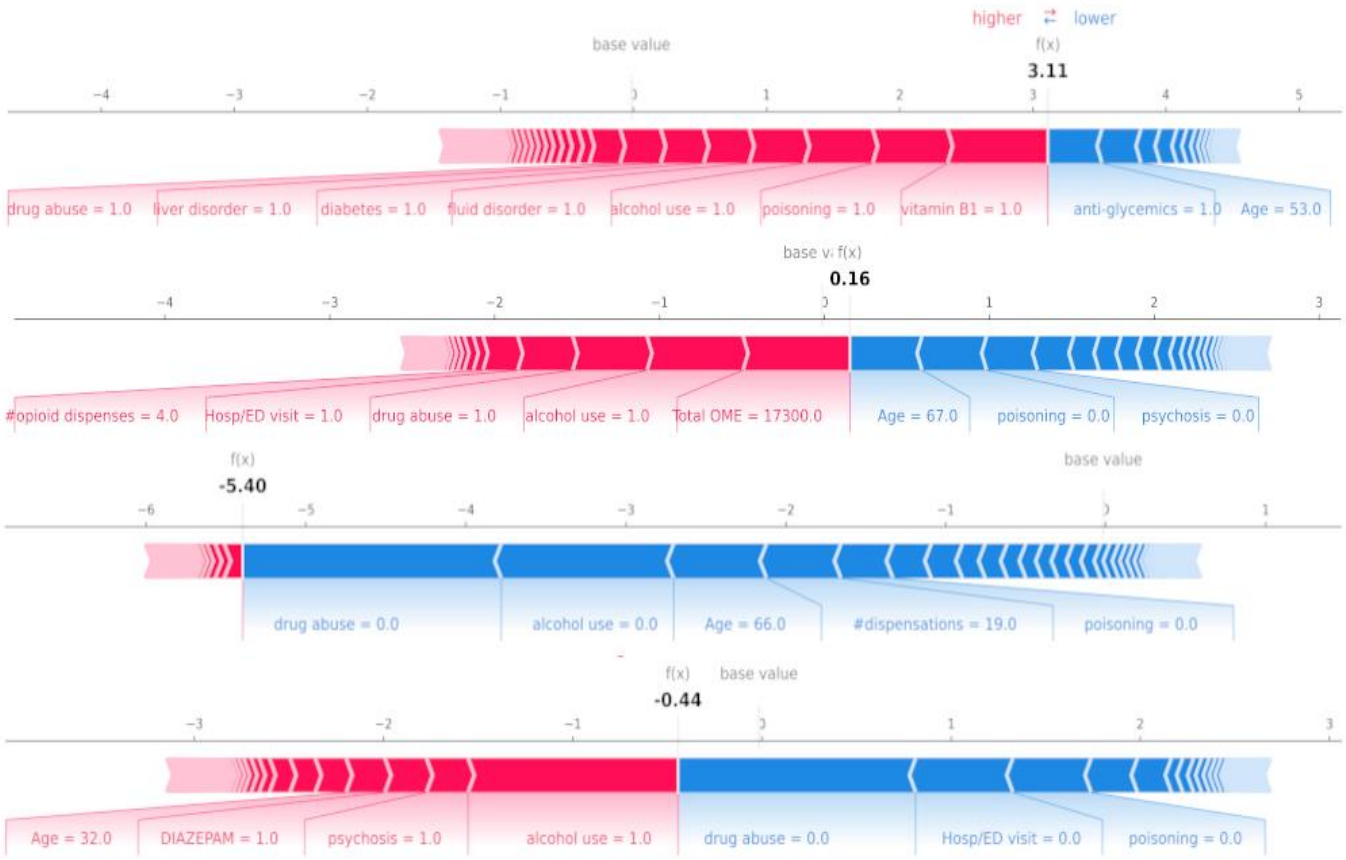

Note: Pharmacy risk score- derived feature using proportion of opioid patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each pharmacy; training and validating the XGBoost classifier with these features alone resulted in an AUC of 0.877 in the 2018 validation set

**(B)**




Note: Pharmacy risk score- derived feature using proportion of opioid/benzodiazepine patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each pharmacy; red indicates higher values of categorical variables and plots to the right of 0.0 indicate the tendency to be associated with the study outcome while blue indicates lower values of categorical variables and plots to the left of 0.0 indicate the tendency to be associated with no outcome

**(C)**

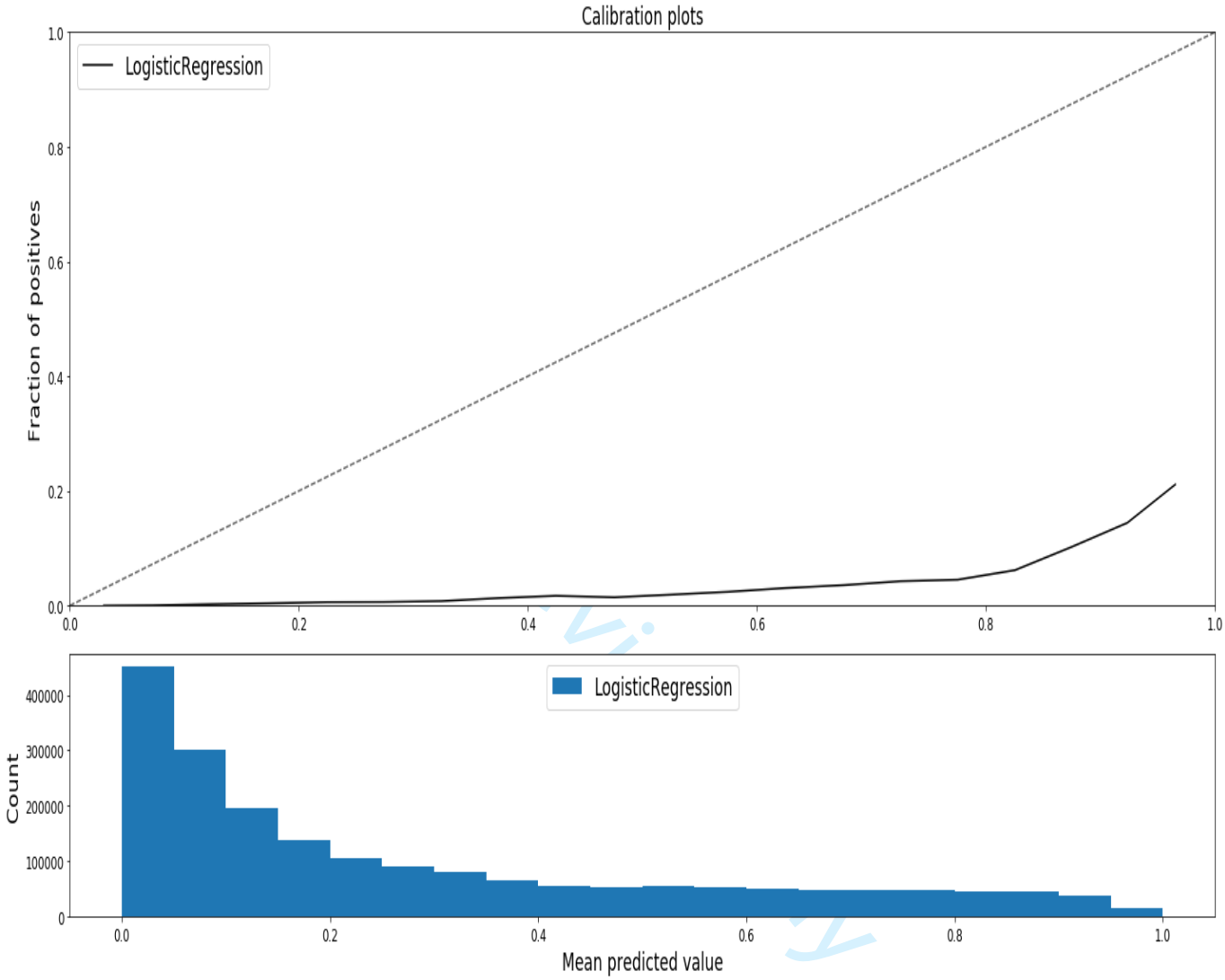



**How to read this figure:** Using hospitalization within 30-days of an opioid dispensation as the outcome of interest, there are 4 scenarios to consider: the XGBoost classifier has low or high confidence in predicting a hospitalization and low or high confidence in predicting **NO** hospitalization. Start at the base SHAP value of near 0.0 ("base value") in which the classifier is not confident in the prediction. SHAP values (in bold) that are above 0.0 indicate a tendency towards a hospitalization while those that are below 0.0 indicate a tendency for **NO** hospitalization. As the SHAP value moves above 0.0, for example 3.11 in the top panel, the classifier's confidence in predicting a hospitalization is higher. As the SHAP value approaches closer to the base value, for example 0.16 in the second panel, the classifier has relatively lower confidence in predicting a hospitalization. When the SHAP value is below 0.0, for example -5.4 in the third panel, the classifier's confidence in predicting **NO** hospitalization is higher and when the SHAP value is closer to 0.0, for example -0.44 in the bottom panel, the classifier has lower confidence in predicting **NO** hospitalization.

The top panel (SHAP value 3.11) depicts an instance predicted to be high risk for our outcome. This individual has a positive history of drug abuse disorder, liver disorder, diabetes, fluid/electrolyte disorder, alcohol use disorder, poisoning and B vitamin use in the prior 30 days. The third panel (SHAP value -5.40) depicts an instance predicted to be low risk (i.e., no hospitalization) and has a negative history for poisoning, drug and alcohol use disorder.

**Note-** drug abuse: drug abuse disorder; poisoning: history of poisoning; vitamin B1: vitamin B1 in prior 30 days; anti-glycemics: anti-glycemic agents in prior 30 days; age: age at opioid dispensation; # opioid dispenses: number of opioid dispensations in prior 30 days; Hosp/ED visit: history of prior hospitalizations and/or emergency visits in past 6 months; Total OME: total oral morphine equivalents in prior 30 days; DIAZEPAM: history of diazepam use in prior 30 days

**eFigure 4.** Calibration curve plotting observed vs. quantiles of estimated risk for the logistic regression (L1) classifier using the 2018 validation dataset. The majority of counts (dispensations) were predicted to be lower risk.



**eReferences**

1. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning, vol. 1. Springer series in statistics New York (2001)
2. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
3. Zhu, H. Zou, S. Rosset, T. Hastie, "Multi-class AdaBoost", 2009.
4. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2011 May 6;2(3):1-27.
5. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
6. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. InProceedings of the 27th international conference on machine learning (ICML-10) 2010 (pp. 807-814).
7. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).

BMJ Open

# Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada

| | |
|---|---|
| Journal: | *BMJ Open* |
| Manuscript ID | bmjopen-2020-043964.R3 |
| Article Type: | Original research |
| Date Submitted by the Author: | 11-May-2021 |
| Complete List of Authors: | Sharma, Vishal; University of Alberta, School of Public Health<br>Kulkarni, Vinaykumar; OKAKI Health Analytics<br>Eurich, Dean; University of Alberta, School of Public Health<br>Kumar, Luke; Alberta Machine Intelligence Institute<br>Samanani, Salim; Okaki Health Intelligence, |
| Primary Subject Heading: | Epidemiology |
| Secondary Subject Heading: | Health informatics, Public health |
| Keywords: | PUBLIC HEALTH, EPIDEMIOLOGY, Adverse events < THERAPEUTICS, Health & safety < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Clinical governance < HEALTH SERVICES ADMINISTRATION & MANAGEMENT |
| | |

*I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.*

*The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.*

*Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.*

**Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada**

**Author list (in order):**

Vishal Sharma (0000-0001-7907-1183), Vinaykumar Kulkarni, Dean T. Eurich (0000-0003-2197-0463), Luke Kumar, Salim Samanani (0000-0001-6751-4805)

**Address for each author:**

2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of Alberta, Edmonton, Alberta, Canada, T6G 2E1 Vishal Sharma BPharm PhD Candidate,

OKAKI Health Intelligence, Edmonton, Alberta, Canada, Vinaykumar Kulkarni MSc

2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of Alberta, Edmonton, Alberta, Canada, T6G 2E1 Dean Eurich professor

Alberta Machine Intelligence Institute, Edmonton, Alberta, Canada, T5J 3B1 Luke Kumar MSc

OKAKI Health Intelligence, Calgary, Alberta, Canada, Salim Samanani MD, Medical Director

**Corresponding Author:**

Dean Eurich, 2-040 Li Ka Shing Center for Health Research Innovation, University of Alberta, Edmonton, Alberta, Canada, T6G 2E1; Phone 780-492-6333; fax 780-492-7455; email: deurich@ualberta.ca

**Acknowledgement**

This study is based on data provided by The Alberta Strategy for Patient Orientated Research (AbSPORU) SUPPORT unit and Alberta Health. The interpretation and conclusions contained herein are those of the researchers and do not necessarily represent the views of the Government of Alberta or AbSPOR. Neither the Government of Alberta, ABSPOR nor Alberta Health expresses any opinion in relation to this study. This work was supported by Mitacs through the Mitacs Accelerate Program (VS and DTE).

**Contributors:** VS VK LK SS and DTE were involved in the conception and design of the study. VS VK LK SS and DTE analyzed the data. VS VK and LK drafted the article. VS VK LK DTE and SS revised the article. All authors gave final approval of the version to be published. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. DTE is the guarantor.

**Funding:** This study received no funding.



***Competing Interest:*** *All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; Salim Samanani has received grants from the College of Physicians & Surgeons of Alberta; no other relationships or activities that could appear to have influenced the submitted work.*

**Ethical approval:** This study was approved by the Health Research Ethics board at the University of Alberta (#Pro00083807_AME1).

**Data Sharing:** The data used in this study is not available for external analysis. However, administrative health data can be accessed from Alberta Health by following defined research protocols and confidentiality agreements.

**Transparency:** The lead author, VS, (the manuscript's guarantor, Dean Eurich) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as originally planned (and, if relevant, registered) have been explained.

**Word Count: 3357**

## Abstract

**Objective:** To develop machine-learning models employing administrative-health data that can estimate risk of adverse outcomes within 30-days of an opioid dispensation for use by health-departments or prescription monitoring programs.

**Design, Setting, and Participants:** This prognostic study was conducted in Alberta, Canada between 2017-2018. Participants included all patients 18 years of age and older who received at least one opioid dispensation. Pregnant and cancer patients were excluded.

**Exposure:** Each opioid dispensation served as an exposure.

**Main Outcomes/Measures:** Opioid related adverse outcomes were identified from linked administrative health-data. Machine-learning algorithms were trained using 2017 data to predict risk of hospitalization, emergency department visit, and mortality within 30-days of an opioid dispensation. Two validation sets, using 2017 and 2018 data, were used to evaluate model performance. Model discrimination and calibration performance were assessed for all patients and those at higher risk. Machine-learning discrimination was compared to current opioid guidelines.

**Results:** Participants in the 2017 training set (n=275,150) and validation set (n=117,829) had similar baseline characteristics. In the 2017 validation set, c-statistics for the XGBoost, logistic regression, and neural-network classifiers were 0.87, 0.87, and 0.80, respectively. In the 2018 validation set (n=393,023), the corresponding c-statistics were 0.88, 0.88, and 0.82. C-statistics from the Canadian guidelines ranged from 0.54-0.69 while the US guidelines ranged from 0.50-0.62. The top 5-percentile of predicted risk for the XGBoost and logistic regression classifiers captured 42% of all events and translated into post-test probabilities of 13.38% and 13.45%, respectively, up from the pre-test probability of 1.6%.

**Conclusion:** Machine-learning classifiers, especially incorporating hospitalization/physician claims data, have better predictive performance compared to guideline or prescription history only approaches when predicting 30-day risk of adverse outcomes. Prescription monitoring programs and health departments with access to administrative data can use machine-learning classifiers to effectively identify those at higher risk compared to current guideline-based approaches.

## Article Summary

### Strengths and Limitations:

- This study incorporated near complete capture of opioid dispensations from community pharmacies and used validated administrative health data.
- This study used commonly available algorithms to train machine-learning models using data which is available to government health departments in all provinces in Canada and other single payer jurisdictions; ML classifiers were evaluated with informative prognostic metrics not usually seen in other studies.
- Our predictive models used dispense events and not medication utilization, which is difficult to capture in administrative data.
- Our training dataset does not account for non-prescription opioids, opioids administered in hospitals, and other risks associated with non-prescription use.

## Introduction

Canada is among the countries with the highest rates of opioid prescribing in the world, making prescription opioid use a key driver of the current opioid crisis[1]; a major part of the policy response to the opioid crisis focuses on endorsing safe, appropriate opioid prescribing[2-4]. In order to minimize high risk opioid prescribing and to identify patients at high risk of opioid related adverse outcomes, numerous health regulatory bodies have released clinical practice recommendations for health providers regarding appropriate opioid prescribing[3,5,6].

Prescription monitoring programs (PMPs) have been implemented around the world, like Alberta's provincial Triplicate Prescription Program (TPP)[7] in Canada, and are mandated to monitor the utilization and appropriate use of opioids to reduce adverse outcomes. In most jurisdictions, both population-level monitoring metrics and clinical decision aids are used to identify patients at risk of hospitalization or death and are most often based on prescribing guidelines. However, a comprehensive infrastructure of administrative data containing patient level International Statistical Classification of Diseases and Related Health Problems (ICD)[8] codes and prescription drug histories exists in Alberta and other provinces in Canada which could be further integrated to predict opioid-related risk. Furthermore, current guidelines addressing high risk prescribing and utilization of opioids were derived from studies that used traditional statistical methods to identify population level risk factors for overdose rather than an individual's absolute risk[3,9,10]; these population estimates may not be generalizable to different populations[11]. Thus, a functional gap exists in many health jurisdictions where much of the available administrative health data is not being leveraged for opioid prescription monitoring.

Supervised machine learning (ML)[12,13] is an approach that uses computer algorithms to build predictive models in the clinical setting that can make use of the large amounts of available administrative data[14,15], all within a well-defined process[16]. Supervised ML trains on labelled data to develop prediction models that are specific to different populations and, in many cases, can provide better predictive performance than traditional, population-based statistical models[10,15,17]. We identified one study[10] that applied ML techniques to predict overdose risk in opioid patients pursuant to a prescription. In their validation sample, they found that the deep neural network (DNN) and gradient boosting machines (GBM) algorithms carried the best discrimination performance based on estimated c-statistics and that the ML approach out-performed the guideline approach in terms of risk prediction; neural networks have little interpretability and are not necessarily better at predicting outcomes when trained on structured data[18]. This study relied on c-statistics to evaluate their ML models and did not emphasize other performance metrics (e.g., positive likelihood ratios, pre and post-test probabilities) required to assess clinical utility that are recommended by medical reporting guidelines[11,13,19,20]. It also did not address the important issue of ML model interpretability[21]. Reporting informative prognostic metrics is needed to better understand the capabilities of ML classifiers if health departments and PMPs are to incorporate them into their decision-making processes.

The objective of our study was to further develop and validate ML algorithms (beyond just DNN) to predict the 30-day risk of hospitalization, emergency visit and mortality for a patient in Alberta, Canada at the time of an opioid dispensation using administrative data routinely available to health departments and PMPs and evaluate them using the above

referenced reporting guidelines. We also analyzed feature importance to provide meaningful interpretations of the ML models. Comparing discrimination performance (area under the receiver operating characteristics curves), we hypothesized that the ML process would perform better than the current guideline approach for predicting risk of adverse outcomes related to opioid prescribing.

## Methods

### Study Design and Participants

This prognostic study used a supervised ML scheme. All patients in Alberta, Canada who received a dispensation for an opioid, were 18 years of age and older between Jan 1, 2017 and Dec 31, 2018 were eligible. Patients were excluded from all analyses if they had any previous diagnosis of cancer, received palliative interventions or were pregnant during the study period (eTable 1 in Supplement) as use of opioids in these contexts is clinically different.

Government health departments and payers in many jurisdictions have systems to capture prescription histories and ICD diagnostic codes. As such, we linked various administrative health data sets available in Alberta, Canada using unique patient identifiers in order to establish a complete description of patient demographics, drug exposures and health outcomes. These databases include *1) Pharmaceutical Information Network (PIN):* PIN data includes all dispensing records from community pharmacies from all prescriber types occurring in the province outside of the hospital setting. PIN collects all drug dispensations irrespective of age or insurance status in Alberta; Anatomical Therapeutic Chemical classification (ATC) codes[22] were used to identify opioid dispensations and their respective opioid molecules (eTable 5), 2*)*

*Population and Vital Statistics Data (VS, Alberta Services)*: sex, age, date of birth, death date, immigration and emigration data, and underlying cause of death according to the World Health Organization algorithm using ICD codes[8], *3) Hospitalizations and Emergency Department Visits (National Ambulatory Care Reporting System [NACRS], Discharge Abstract Database [DAD]):* all services, length of stay, diagnosis (up to 25 ICD-10[8] based diagnoses). Data and coding accuracy are routinely validated both provincially and centrally via the Canadian Institute for Health Information, and 4*) Physician Visits/Claims (Alberta Health):* all claims from all settings (e.g., outpatient, office visits, emergency departments, inpatient) with associated date of service, ICD code, procedure and billing information.

This study followed the TRIPOD and STARD reporting guidelines[23-25] and received ethics approval from the University of Alberta ethics board (Pro00083807_AME1).

**Measures and Outcome**

ML models were trained on a labelled dataset in which the observation/analysis unit was an opioid dispensation. Every opioid dispensation, not just the incident one, was used as a potential instance to predict the risk of our outcome. The primary outcome was a composite of a drug-related hospitalization, emergency department (ED) visit or mortality within 30 days of an opioid dispensation based on ICD-10 codes used by others and identified from DAD, NACRS and Vital Statistics (T40, F55, F10-19; eTable 2 in Supplement)[2,10,26].

We anticipated that our defined outcome would be a rare event, leading to a class imbalanced dataset[27]. To address this, we relied on specifying balanced class weightage for supporting algorithms; other approaches were deemed not suitable (e.g., oversampling using

randomly repeating minority class); under sampling (sub-sampling within the majority class) resulted in changes in outcome prevalence. Class weightage is a commonly used method[28] to address class imbalance along with over and under-sampling approaches. However, oversampling, which involves generating new opioid dispensations from the original data distribution and is prone to introducing bias, is difficult due to the categorical nature of the data and beyond the scope of this study. With under-sampling, which takes samples from the majority class (in this case, no 30-day event after dispensation), we would not be able to use all of the information provided by the data in instances with no outcome. Hence, we decided to use the class weightage method which does not alter the data distribution. Instead, the learning process is adjusted in a way that increases the importance of the positive class (instances that led to a 30-day event)[29].

**Predictor Candidates for ML Models**

Predictor variables in our ML models included those that were informed by the literature[3,4,10] and those directly obtained from the data sets. These included features based on demographics (age, sex, income using Forward Sortation index from postal codes[30]), co-morbidity history using ICD-based Elixhauser score categories[31], health care utilization (number of unique providers, number of hospital and emergency department visits), and drug utilization (level 3 ATC codes[22], oral morphine equivalents[32], concurrent use with benzodiazepines, number of opioid and benzodiazepine dispensations, number of unique opioid and benzodiazepine molecules). Depending on the potential predictor and data availability, we used data from 30 days to 5 years before the opioid dispensation to generate model features (eFigure 1 in Supplement); 30 days was used to reflect the immediate nature of the risk and 5

years to fully capture co-morbidities. This approach aligns with how health providers would assess patients using the entire history of co-morbidities and then the more immediate factors in deciding on the need for a therapeutic as well as risk in patients. We performed experiments to identify the features and data sets that contributed most to predicting the outcomes with a view to minimizing the potential future data requirements for health departments and PMPs.

**Statistical Analyses and Machine-Learning Prediction Evaluation**

We randomly divided the patients in the 2017 portion of our study cohort into training (70%) and validation (30%) sets[13] by patients and opioid dispensations such that no patients in the training set were in the validation set. Baseline characteristics and event rates were compared in the training vs validation group, and between those who experienced the outcome and those who did not using chi-squared tests of independence. As well, we used all the 2018 data as another independent validation set.

We trained commonly used[13,33] ML algorithms (eAppendix in Supplement) and further tuned out-of-box models using 5-fold cross validation on the training data to address model overfitting[13,34]. As is common in ML validation studies[10,13], we reported model discrimination performance (i.e. how well a model differentiates those at higher risk from those at lower risk)[11] using area under the receiver operating characteristic curve (AUROC; c-statistic). We then stratified the two ML models with the highest c-statistics into percentile categories (deciles) according to absolute risk of our outcome, as was done in previous studies[10,35]. We also plotted AUROC[11] and precision-recall curves (PRCs)[36].

Because discrimination alone is insufficient to assess ML model prediction capability, we assessed a second necessary property, namely, calibration (i.e., how similar the predicted absolute risk is to the observed risk across different risk strata)[11,37]. Using the two ML models with the highest discrimination performance, we assessed calibration performance on the 2018 data by plotting observed (fraction of positives) vs predicted risk (mean predicted value). Using these same two ML classifiers, we analyzed the top 0.1, 1, 5, and 10 percentiles of predicted risk by the number of true and false positives, positive likelihood ratios (PLR)[20], positive predictive values (PPV), post-test probabilities, and number needed to screen. We also performed a simulation of daily data uploads for 2018 Quarter 1 to view the predictive capabilities if a ML risk predictor were to be deployed into a monitoring workflow.

For the XGBoost and logistic regression classifiers, we reported feature importance[33] and plotted PRCs that compared all dispenses to those within the top 10 percentiles of estimated risk. As well, for the XGBoost classifier, we described feature importance on model outcome using SHAP values[38,39] to add an additional layer of interpretability.

Finally, we compared ML risk prediction (the two ML models with highest discrimination performance) to current guideline approaches as others have[10], using the 2019 Centers for Medicare & Medicaid Services (CMS) opioid safety measures[40] and the 2017 Canadian Opioid Prescribing Guideline[3]. This was done by using the guidelines as "rules" when coding for the 30-day risk of event at the time of each opioid dispensation on the entire 2018 validation set. We also compared the discrimination performance of different logistic regression classifier models using various combinations of features derived from their respective databases: **1)**

demographic and drug/health utilization features from PIN and 2**)** co-morbidity features derived from DAD, NACRS and Claims.

All analyses were done using Python (v. 3.6.8,), SciKit Learn[41] (v. 0.23.2) SHAP[39] (v. 0.35), XGBoost (v. 0.90)[42], Pandas (v. 1.0.5)[43] and H20 Driverless AI (version 1.9).

**Patient and Public Involvement**

This research was done without patient involvement. Patients were not invited to comment on the study design and were not consulted to develop patient relevant outcomes or interpret the results. Patients were not invited to contribute to the writing or editing of this document for readability or accuracy. There are no plans to disseminate the results of the research to study participants.

**Results**

**Patient Characteristics and Predictors**

We identified 392,979 patients with at least one opioid dispensation in 2017 (Figure 1). This cohort was used to train (n= 275,150, 70%) and validate (n=117,829, 30%) ML models. In 2017 and 2018, 6,608 and 5,423 patients experienced the defined outcome, respectively. Baseline characteristics were different between those who experienced the outcome and those who did not (eTable 3 in Supplement) while characteristics were similar between the training and validation sets (eTable 4 in Supplement). There were 2,283,075 opioid dispensations in 2017 and 1,977,389 in 2018. Overall, in 2017, 2.03% (n= 45,757) of opioid dispensations were associated with the outcome; in 2018, the estimate was 1.6% (n= 31,392).

As described above, we categorized our candidate features into four groups (eTable 5 in Supplement). When using all of the databases, the total number of features was 283 and 34 when considering only co-morbidities.

**Machine-Learning Prediction Performance**

Using the 2017 validation set, AUROCs for the XGBoost and logistic regression classifiers had the highest discrimination performance at 0.87, while the neural network classifier had lower performance at 0.80 (eTable 6 in Supplement).

Discrimination performance was similar for the 2018 validation set (n=393,023; eTable 6 in Supplement). XGBoost and logistic regression had the highest estimated AUROCs and area under PRCs while the neural network classifier was lower (Figure 2A, 2B). As expected, precision-recall curves indicate stronger predictive performance in opioid dispensations at higher predicted risk percentiles (Figure 2C, 2D).

In the 2018 validation set, although discrimination performance was similar (0.88), individual feature importance was different between the logistic regression and XGBoost classifiers, with logistic regression feature importance more reliant on co-morbidity data from DAD, NACRS and Claims while XGBoost relied more on drug utilization data from PIN (eFigure 2). With the XGBoost classifier, history of drug abuse, alcoholism, and prior hospitalization/emergency visit carried the highest importance for predicting the study outcome (eFigure 3A) where the presence of these features in a patient suggested a strong prediction towards having the defined outcome (eFigure 3B and 3C).

**Calibration**

When considering dispensations predicted to be in the highest percentiles of risk, the top 5-percentile captured 42% of all outcomes using the XGBoost and logistic regression classifiers (Table 1). Also, as the predicted risk percentiles get higher (top 10 percentile to top 0.1 percentile), so too do the corresponding PPVs with the top 0.1 percentile associated with a PPV of 33% for the XGBoost classifier. As well, lower categories of risk percentiles were associated with lower outcomes (Figure 3, eFigure 4). When we simulated a monitoring workflow scenario with daily data uploads, a similar pattern was illustrated where the dispensations predicted to be higher risk had higher event rates (Figure 4).

After using the XGBoost and logistic regression classifiers to identify the dispensations in the highest predicted risk percentiles, the pre-test probability of the outcome (1.6%) was transformed into higher post-test probabilities, with higher probabilities in the riskier percentiles (Table 1). The number needed to screen also decreased as predicted risk increased (Table 1).

Comparing discrimination performance, ML risk prediction outperformed the current guideline approaches when using various combinations of guideline recommendations (Table 2). In many of the guideline scenarios, the estimated AUROCs were close to the 0.5 mark. When we estimated the discrimination performance of the logistic regression classifier based on database source, using all databases produced an AUROC of 0.88. Reducing the database source to only DAD, NACRS, Claims (co-morbidities only) resulted in an AUROC of 0.85, while PIN (prescription history) only was 0.78 (Table 3).

**Discussion**

This study showed that ML techniques using available administrative data (prescription histories and ICD codes) may provide enough discriminatory performance to predict adverse outcomes associated with opioid prescribing. Indeed, our ML analyses showed very high discrimination performance at 0.88. The linear model (logistic regression) and XGBoost carried higher discrimination and calibration performance, while the neural network classifier did not perform as well. By identifying the predicted top 5-10 percentile of absolute risk pursuant to an opioid dispensation, we were able to capture approximately half of all outcomes using ML methods. All ML models we trained had higher discrimination performance using the validation sets compared to the clinical guideline approach.

Since the prevalence of our defined outcome is relatively low in the general population, PPVs would also be expectedly low. However, estimated PPVs increased when we considered higher risk dispensations, as is expected since PPV is related to event prevalence. This is important because different users of a risk predictor will require different predictive capabilities. Similarly, our estimates of positive likelihood ratios and associated post-test probabilities also increased in dispensations with higher predicted risk indicating the strong predictive capabilities of the XGBoost and logistic regression classifiers; likelihood ratios >10 generate conclusive changes from pre-test to post-test probabilities[20].

The current guideline approach to assess absolute opioid prescribing risk produced c-statistic estimates closer to 0.5 indicating that discrimination was not much better than chance alone. ML models with higher predictive performance can better support health departments and PMPs with monitoring mandates to identify and intervene on those at high risk and their associated prescribers. We also found that adding co-morbidity features from administrative

databases increased prediction performance compared to prescription history alone, thus making the case for the use of this data by PMPs and health departments. However, if only prescription history is available, our trained XGBoost classifier still had strong discrimination performance.

We found only one study that used ML approaches to quantify the absolute risk of an event pursuant to an opioid dispensation[10]. Their methodology used rolling 3-month windows for estimating risk and ML model training while we used historic records to estimate 30-day risk. Differences in study population and feature selection may explain why their highest performing ML model was deep learning (neural network classifier) and ours was not. Nevertheless, we were able to replicate their predictive performance using our ML approach as we both showed that ML approaches have higher predictive capabilities than guideline approaches. Both of our studies used predicted percentile risk estimates to identify high risk dispensations and were able to do so with strong discrimination and calibration performance. Furthermore, we emphasized prognostic metrics which are more informative to assess the clinical utility of ML classifiers using pre- and post-test probabilities, something not done in other studies and recommended in medical guidelines[20]. This major aspect of our study, not done previously, is important because any ML classifier that does not increase prognostic information compared to baseline cannot be incorporated into decision making for the purpose of intervening on higher risk instead of lower risk patients. Indeed, another study we found describes how identifying cases in higher predicted risk percentiles using ML methods can be deployed in hospital settings for the purpose of targeted interventions[35] upon discharge, however the effect on outcomes is still to be determined.

The limitations of our study are similar to other ML studies[10] and need to be addressed when considering deployment of ML risk predictors. Our training dataset was not able to account for non-prescription opioid consumption and the risk associated with non-prescription use, both of which are substantial contributors to overall risk[2]. Regarding our analysis, we assumed that all dispensations were independent events; future research in this area should focus on employing ML methods using correlated data. As with all ML projects, our models were trained using Alberta data and might not be generalizable to other populations, or to specific populations within Alberta. However, one of the benefits of the ML process is that models can be retrained or similar methods could be used to develop new models to accommodate different populations.

This study suggests that ML risk prediction can support PMPs, especially if readily available administrative health data is used. PMPs currently use population-based guidelines which we, and others, have shown cannot predict absolute individual risk. The ML process allows for flexibility in model training, validation and deployment to specific settings in which, for the case of PMPs, high risk patients can be identified and targeted for intervention either at the patient or provider level. For example, a ML classifier can be trained on accessible data to create an aggregated list of "high risk" patients at regular time intervals to identify points of intervention. Moreover, ML classifiers can be retrained over time as changes in populations and trends in prescribing occur and are therefore specific to the population unlike broadly based guidelines. Further research can assess whether implementation of a ML-based monitoring system by PMPs leads to improved clinical outcomes within their own jurisdictions

and whether other available features or feature reduction can yield sufficiently valid results for their own intended purposes.

## References


1. Belzak L, Halverson J. Evidence synthesis - The opioid crisis in Canada: a national perspective. *Health Promotion and Chronic Disease Prevention in Canada.* 2018;38(6):224-233.
2. Gomes T, Khuu W, Martins D, et al. Contributions of prescribed and non-prescribed opioids to opioid related deaths: population based cohort study in Ontario, Canada. *BMJ.* 2018;362:k3207.
3. Busse JW, Craigie S, Juurlink DN, et al. Guideline for opioid therapy and chronic noncancer pain. *Canadian Medical Association Journal.* 2017;189(18):E659-E666.
4. Dowell D. CDC guideline for prescribing opioids for chronic pain. 2016.
5. ismp Canada. Essential Clinical Skills for Opioid Prescribers. 2017; https://www.ismp-canada.org/download/OpioidStewardship/Opioid-Prescribing-Skills.pdf. Accessed Nov 2018.
6. Centre for Effective Practice. Management of Chronic Non Cancer Pain. 2017; thewellhealth.ca/cncp.
7. College of Physicians and Surgeons of Alberta. TPP Alberta – OME and DDD Conversion Factors. 2020; http://www.cpsa.ca/tpp/. Accessed Jun 2020.
8. World health Organization. Classification of Diseases (ICD). 2019; https://www.who.int/classifications/icd/icdonlineversions/en/. Accessed Jun 2020.
9. Gomes T, Mamdani MM, Dhalla IA, Paterson JM, Juurlink DN. Opioid Dose and Drug-Related Mortality in Patients With Nonmalignant PainOpioid Dose and Drug-related Mortality. *JAMA Internal Medicine.* 2011;171(7):686-691.
10. Lo-Ciganic W-H, Huang JL, Zhang HH, et al. Evaluation of Machine-Learning Algorithms for Predicting Opioid Overdose Risk Among Medicare Beneficiaries With Opioid Prescriptions. *JAMA network open.* 2019;2(3):e190968-e190968.
11. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. *JAMA.* 2017;318(14):1377-1384.
12. Shah NH, Milstein A, Bagley P, Steven C. Making Machine Learning Models Clinically Useful. *JAMA.* 2019;322(14):1351-1352.
13. Liu Y, Chen P-HC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA.* 2019;322(18):1806-1816.
14. Bastanlar Y, Ozuysal M. Introduction to machine learning. *Methods in molecular biology (Clifton, NJ).* 2014;1107:105-128.
15. Thottakkara P, Ozrazgat-Baslanti T, Hupf BB, et al. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PloS one.* 2016;11(5):e0155705.
16. Alberta Machine Intelligence Institute. Machine Learning Process Lifecycle. In:2019.
17. Hsich E, Gorodeski EZ, Blackstone EH, Ishwaran H, Lauer MS. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circulation: Cardiovascular Quality and Outcomes.* 2011;4(1):39-45.
18. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Paper presented at: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining2015.
19. Yusuf M, Atal I, Li J, et al. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. *BMJ open.* 2020;10(3):e034568.
20. Jaeschke R, Guyatt GH, Sackett DL, et al. Users' Guides to the Medical Literature: III. How to Use an Article About a Diagnostic Test B. What Are the Results and Will They Help Me in Caring for My Patients? *JAMA.* 1994;271(9):703-707.
21. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA.* 2018;320(21):2199-2200.

22. World Health Organization. International language for drug utilization research, ATC/DDD. 2020; https://www.whocc.no/. Accessed Jun 2020, 2020.
23. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine.* 2015;162(1):W1-W73.
24. equator network. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. 2020; https://www.equator-network.org/reporting-guidelines/tripod-statement/. Accessed Feb 2020.
25. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open.* 2016;6(11):e012799.
26. Zhou H, Della PR, Roberts P, Goh L, Dhaliwal SS. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. *BMJ Open.* 2016;6(6):e011060.
27. Brownlee J. A Gentle Introduction to Imbalanced Classification. 2020; https://machinelearningmastery.com/what-is-imbalanced-classification/. Accessed Jan 2021.
28. King G, Zeng L. Logistic regression in rare events data. *Political analysis.* 2001;9(2):137-163.
29. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. *Journal of Big Data.* 2019;6(1):1-54.
30. Government of Canada. Forward Sortation Area—Definition. 2015; https://www.ic.gc.ca/eic/site/bsf-osb.nsf/eng/br03396.html. Accessed April 2020, 2020.
31. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical care.* 2005:1130-1139.
32. College of Physicians and Surgeons of Alberta. OME and DDD conversion factors. http://www.cpsa.ca/wp-content/uploads/2017/06/OME-and-DDD-Conversion-Factors.pdf.
33. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European heart journal.* 2017;38(23):1805-1814.
34. Rose S. Machine Learning for Prediction in Electronic Health Data. *JAMA Network Open.* 2018;1(4):e181404-e181404.
35. Morgan DJ, Bame B, Zimand P, et al. Assessment of Machine Learning vs Standard Prediction Rules for Predicting Hospital Readmissions. *JAMA Network Open.* 2019;2(3):e190348-e190348.
36. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE.* 2015;10(3):e0118432.
37. Shah ND, Steyerberg EW, Kent DM. Big Data and Predictive Analytics: Recalibrating Expectations. *JAMA.* 2018;320(1):27-28.
38. Molnar C. *Interpretable machine learning. A Guide for Making Black Box Models Explainable.* 2019.
39. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Paper presented at: Advances in neural information processing systems2017.
40. Centers for Medicare & Medicaid Services (CMS). Announcement of calendar year (CY) 2019 Medicare Advantage capitation rates and Medicare Advantage and Part D payment policies and final call letter.
41. Buitinck L, Louppe G, Blondel M, et al. API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:13090238.* 2013.
42. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Paper presented at: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining2016.
43. The pandas development team. pandas-dev/pandas: Pandas. 2020; https://doi.org/10.5281/zenodo.3509134, Jan 2021.

**Figure Legend**

Figure 1. Patient flow diagram of study participants used for training and validating ML models. NACRS: National Ambulatory Care Reporting System; DAD: Discharge Abstract Database; VS: Vital Statistics; PIN: Pharmaceutical Information Network; Claims: Physician Claims

Figure 2. Area under the receiver operating characteristic curve (AUROC) (A) and precision-recall curves (B) for all dispensations using logistic regression (L1), neural network, support vector machine (SVM), XGBoost and Naïve-Bayes; precision-recall curves for higher risk dispensations according to predicted risk percentile categories for logistic regression (C) and XGBoost (D) using the 2018 validation set.

Figure 3. Calibration curve plotting observed vs. quantiles (deciles) of estimated risk for the XGBoost classifier using the 2018 validation dataset. The majority of counts (dispensations) were predicted to be lower risk.

Figure 4. Simulation of a clinical workflow with daily uploads and events per 100 daily dispenses by risk percentiles using 2018 Quarter 1 (Q1) data for logistic regression (A) and XGBoost (B) classifiers.

Table 1. Highest percentiles of estimated risk and predictive performance using the XGBoost and logistic regression classifiers for the 2018 validation dataset (n=393,023). Total number of dispenses= 1,977,389; total number of outcomes= 31,392.

| Metric | Top 0.1%ile | | Top 1%ile | | Top 5%ile | | Top 10%ile | |
|---|---|---|---|---|---|---|---|---|
| | XGBoost | Logistic Regression | XGBoost | Logistic Regression | XGBoost | Logistic Regression | XGBoost | Logistic Regression |
| **Number of Dispenses** | 1,977 | 1,977 | 19,774 | 19,774 | 98,869 | 98,869 | 197,739 | 197,739 |
| **TP captured** | 655 | 472 | 4204 | 4100 | 13224 | 13293 | 18404 | 18409 |
| **Percent of TP** | 2.09 | 1.50 | 13.39 | 13.06 | 42.13 | 42.35 | 58.63 | 58.64 |
| **FP captured** | 1322 | 1505 | 15570 | 15674 | 85645 | 85576 | 179335 | 179330 |
| **PPV** | 33.13 | 23.87 | 21.26 | 20.73 | 13.38 | 13.45 | 9.31 | 9.31 |
| **PLR** | 30.71 | 19.44 | 16.74 | 16.22 | 9.57 | 9.63 | 6.36 | 6.36 |
| **Post-test Probability*** | 33.13 | 23.87 | 21.26 | 20.73 | 13.38 | 13.45 | 9.31 | 9.31 |
| **NNS** | 3.17 | 4.49 | 5.08 | 5.22 | 8.48 | 8.43 | 12.95 | 12.95 |

*Pre-test probability estimated at 1.6% using prevalence.

TP: true positives; FP: false positives; PPV: positive predictive value; PLR: positive likelihood ratio; NNS: number needed to screen

Note: Logistic regression used L1 (lasso) parameter regularization

Table 2. Discrimination performance of guideline approach using the 2018 validation set. Guideline approaches were adapted from the 2017 Canadian Opioid Prescribing Guideline and 2019 Centers for Medicare & Medicaid Services (CMS) opioid safety measures and compared to logistic regression and XGBoost classifiers (each with an estimated area under the receiver operating characteristic curve of 0.88). These guidelines were used as rules to predict the 30-day risk of event at the time of opioid dispensation.

| **Canadian Guidelines *** | **AUROC** | **Sensitivity** | **Specificity** |
|---|---|---|---|
| History of mental disorder only | 0.620 | 0.90 | 0.34 |
| Substance abuse only | 0.686 | 0.99 | 0.37 |
| OME/day >90 only | 0.539 | 0.22 | 0.85 |
| (Mental disorder and substance abuse) OR OME/day >90 | 0.690 | 0.91 | 0.47 |
| Mental disorder and substance abuse AND OME/day >90 | 0.560 | 0.20 | 0.91 |
| Mental disorder OR substance abuse OR OME/day >90 | 0.589 | 0.99 | 0.18 |
| **CMS Guidelines**** | | | |
| High opioid dose (>120 OME/day for 90+days) | 0.507 | 0.081 | 0.933 |
| Concurrency (Opioid & BZRA for 30+ days) | 0.575 | 0.423 | 0.727 |
| Multiple doctors (>4) | 0.591 | 0.294 | 0.888 |
| Multiple pharmacies (>4) | 0.537 | 0.120 | 0.959 |
| All conditions | 0.50 | 0.001 | 0.999 |
| Any condition | 0.622 | 0.62 | 0.625 |

OME: daily oral morphine equivalents; BZRA: benzodiazepine receptor agonist. Elixhauser scoring ICD codes were used to identify mental disorders and substance abuse.

*The Canadian guidelines do not specify timelines. >90 OME was determined by taking the average daily OME over the 30 days prior to dispensation

**The CMS guidelines specify 90 or more days at >120 OME and concurrent use of opioids and benzodiazepines for 30 days or more within an assessment period of 180 days.

Table 3. Discrimination performance based on database source using area under the receiver operating characteristic curve (AUROC) for the logistic regression classifier on the 2018 validation set.

| **Database source** | **Predictor Variables formed from database** | **AUROC** | **Number of features** |
|---|---|---|---|
| PIN only | Drug utilization + Prescription history | 0.78 | 248* |
| DAD, NACRS, Claims | Co-morbidities | 0.85 | 34 |
| PIN, DAD NACRS, Claims (all databases used in study) | Demographic + Drug Utilization + Healthcare Utilization + Co-morbidities | 0.88 | 283 |

Note: drug utilization includes features describing oral morphine equivalents[32], concurrent use with benzodiazepines, number of opioid and benzodiazepine dispensations, number of unique opioid and benzodiazepine molecules; health care utilization includes features describing number of unique health providers visited, number of hospital/emergency department visits; logistic regression used L1 (lasso) parameter regularization; PIN- Pharmaceutical Information Network; DAD- Discharge Abstract Database; NACRS- National Ambulatory Care Reporting System

Figure 1. Patient flow diagram of study participants used for training and validating ML models. NACRS: National Ambulatory Care Reporting System; DAD: Discharge Abstract Database; VS: Vital Statistics; PIN: Pharmaceutical Information Network; Claims: Physician Claims

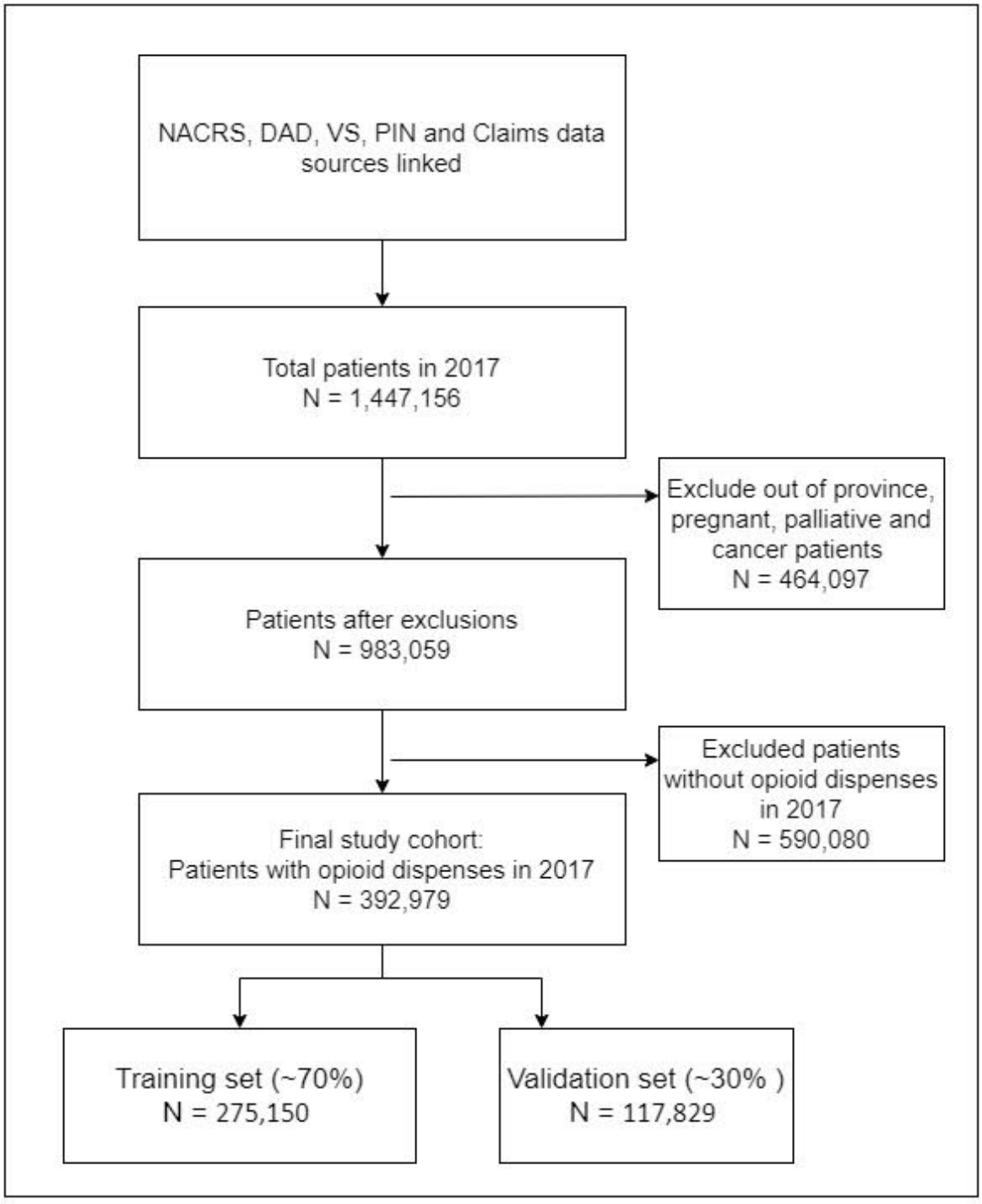
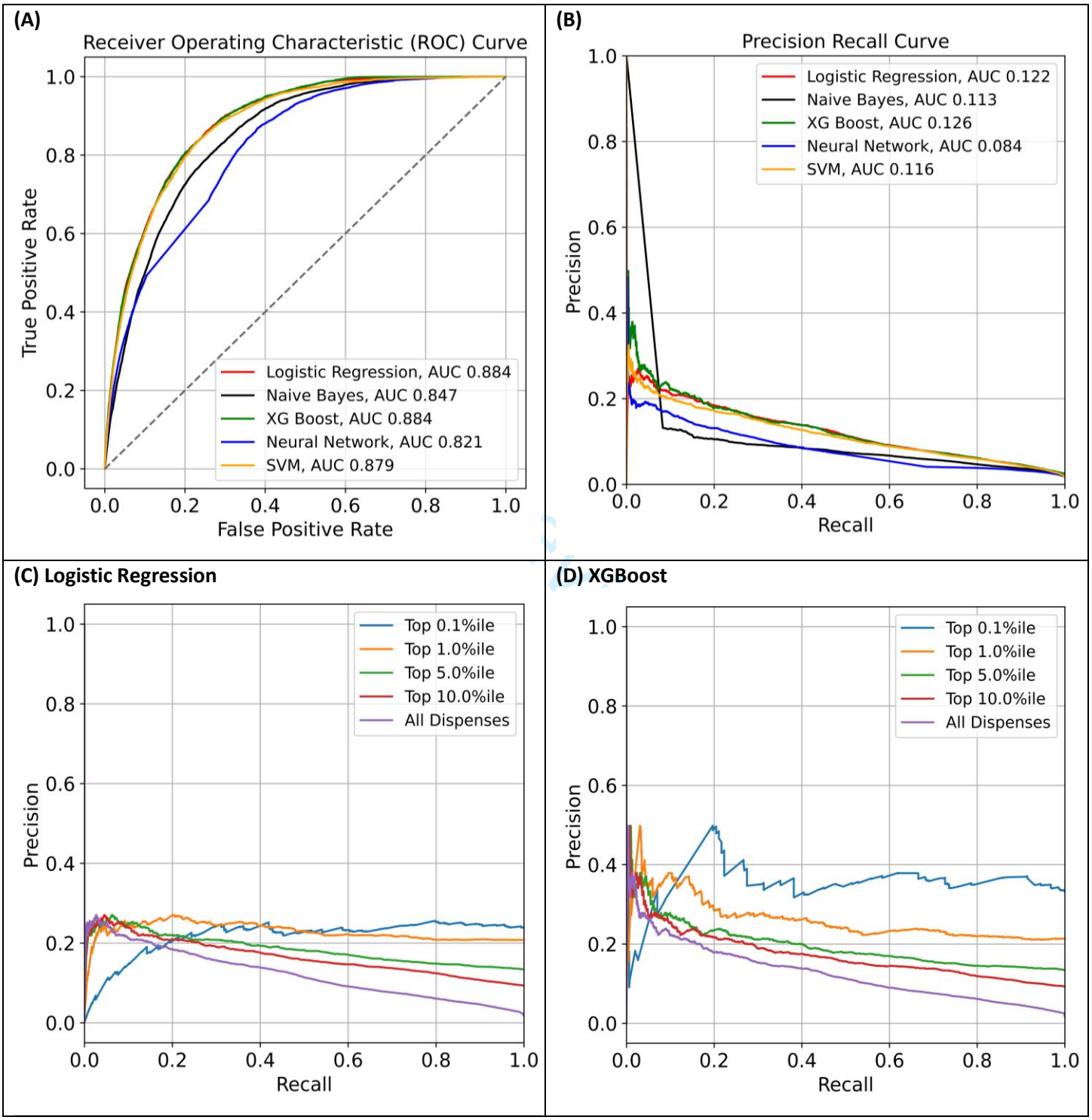


Figure 2. Area under the receiver operating characteristic curves (A) and precision-recall curves (B) for all dispensations using logistic regression (L1), neural network, support vector machine (SVM), XGBoost and Naïve-Bayes; precision-recall curves for higher risk dispensations according to predicted risk percentile categories for logistic regression (C) and XGBoost (D) using the 2018 validation set.




AUC: area under the curve

Figure 3. Calibration curve plotting observed vs. quantiles (deciles) of estimated risk for the XGBoost classifier using the 2018 validation dataset. The majority of counts (dispensations) were predicted to be lower risk.

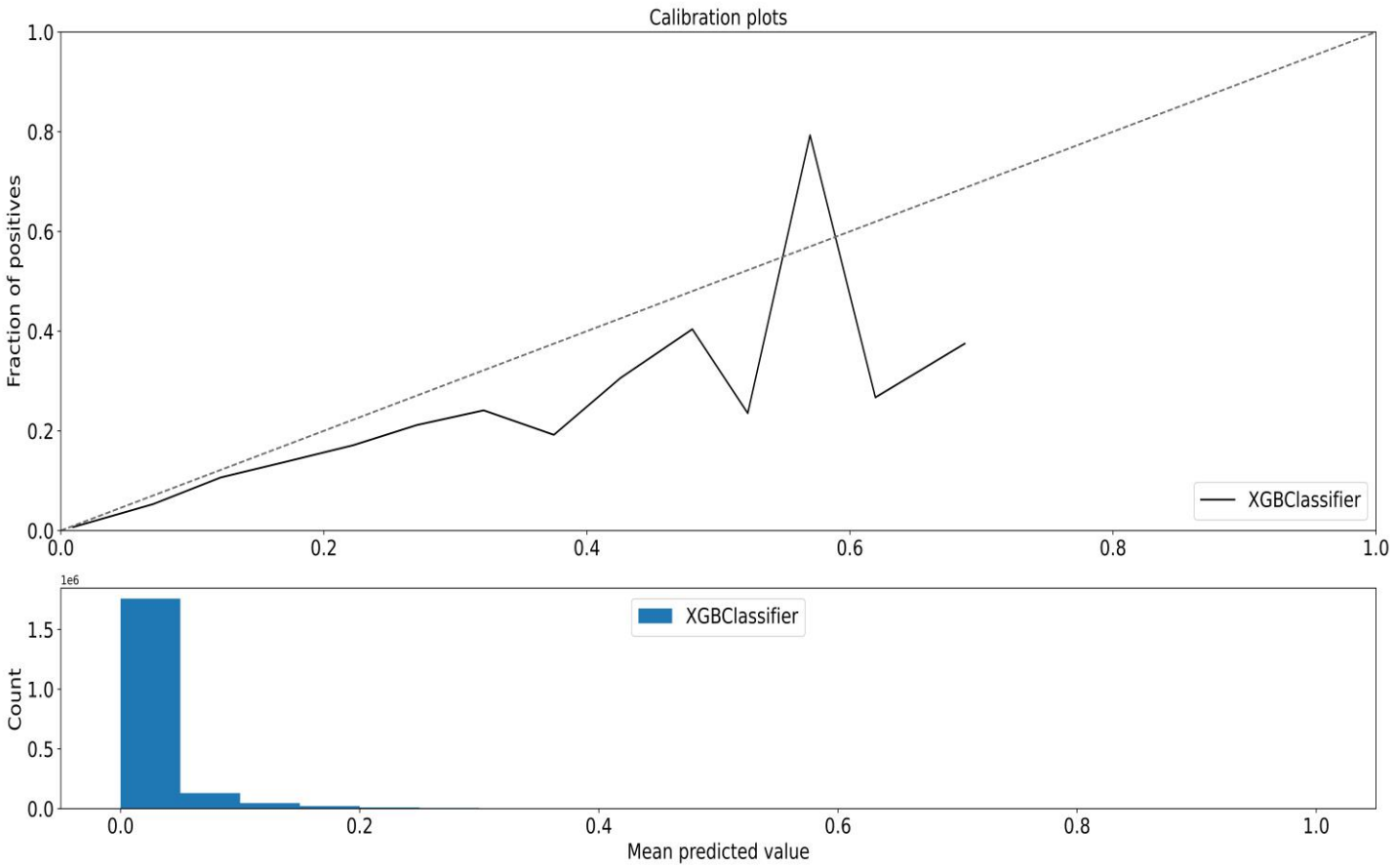
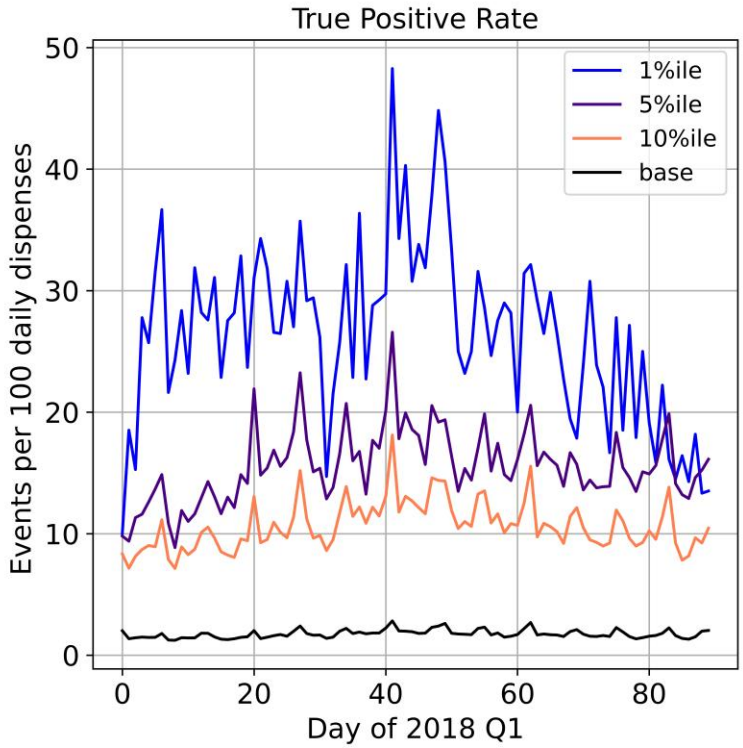
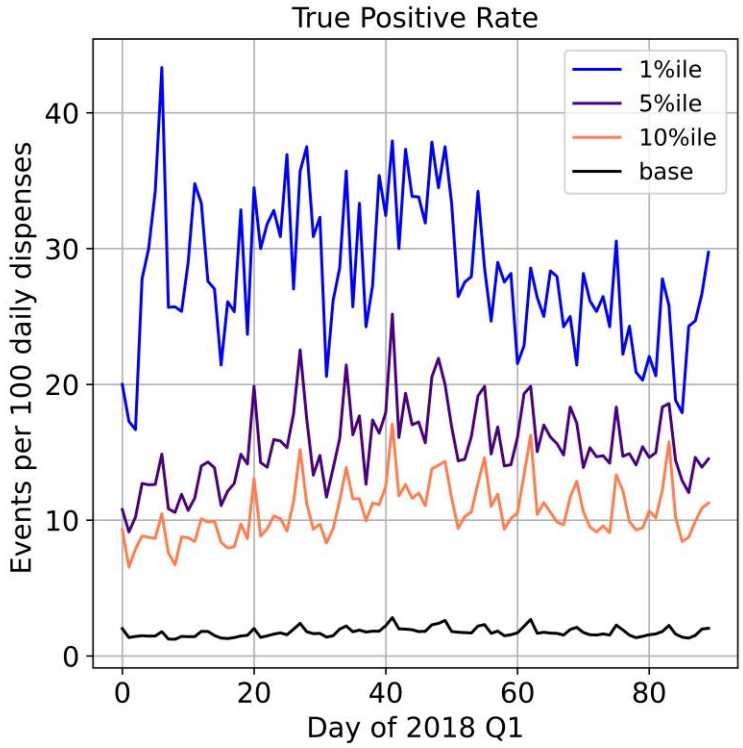


Figure 4. Simulation of a clinical workflow with daily uploads and events per 100 daily dispenses by risk percentiles using 2018 Quarter 1 (Q1) data for logistic regression (A) and XGBoost (B) classifiers.

(A) Logistic Regression (L1)




(B) XGBoost



**Supplementary Content**

**eAppendix.** Machine learning algorithms

**eTable 1.** Diagnostic codes used to exclude patients who had cancer, were pregnant, or were under palliative care.

**eTable 2.** Diagnostic codes used to identify the defined study outcome from emergency visit, hospitalization and death data.

**eTable 3.** Baseline characteristics of study patients (n=392,979). Co-morbidities were determined using Elixhauser criteria. All p-values in the $chi^2$ test of independence were <0.001 unless otherwise indicated.

**eTable 4.** Characteristics of study participants between training and validation groups using 2017 data.

**eTable 5.** Candidate predictors used to train ML algorithms.

**eTable 6.** Discrimination performance using area under the receiver operating characteristic curve (AUROC) of various ML algorithms. Training and validation were done using 2017 data (n=393,979); another independent validation was performed using 2018 data (n=393,023).

**eFigure1.** Schematic of study design and feature generation

**eFigure2**. Feature importance from logistic regression and tree-based (XGBoost) classifiers using the 2018 validation set.

**eFigure3**. Shapley values and feature impact in the XGBoost classifier using the 2018 validation set to describe "associations" between features and the outcome.

**eFigure 4.** Calibration curve plotting observed vs. quantiles of estimated risk for the logistic regression (L1) classifier using the 2018 validation dataset. The majority of counts (dispensations) were predicted to be lower risk.

**eReferences.**

# eAppendix. Machine Learning Algorithms

## Introduction

While there are always updates and new methods coming up in the fields of machine learning, in this study, we have focused on some of the most reliable and proven approaches for predictive modelling which are explainable and popularly used in previous studies of similar nature.

## Logistic Regression

Regression analysis models the relationship between a dependent variable and a set of independent variables [1]. Typically, this includes understanding how the value of the dependent variable changes with the changes in the values of independent variables. Logistic regression [1] uses the logistic function to model a binary dependent variable, where, based on the values of the independent variables the model can approximate one of the two classes, the instance belongs to. This basic binary model can be extended to deal with multiple classes (e.g. One-vs-all classifiers). However, logistic regression is only capable of modeling a linear relationship of independent variables to the dependent variable, hence limited to problems with linear decision boundaries. We used the sci-kit learn library in our experiments[6] and found L1 regularization to be more effective.

## Ridge Classifier

We used the ridge classifier implemented in the Scikit learn library[5]. It implements a classifier using ridge regression which uses an L2 regularization on the least square objective function. The library converts the labels into -1 and 1 and fits a linear regression on the converted labels with the regularization.

## Random Forest

Random forest is a tree ensemble learning algorithm that has wide applicability in many domains[1]. Random forest is a nonlinear learning algorithm, which arrives at nonlinear decision boundaries by independently combining multiple decision trees. Each individual decision tree in the forest can be grown independently of each other on a subset of the training data. Random forests are mainly sensitive to the number of trees, the depth of a tree and the number of covariates randomly chosen to split at each node[1]. These hyper-parameters can be tuned to find the best configuration of every dataset. Random Forests, in general, are less prone to overfit since they always grow individual trees on a subset of the training data[1]. At prediction time, the decision of each tree is aggregated to compute the final prediction.

## Neural Networks (NN)

Neural networks are another collection of non-linear learning algorithms with high representation power. They are known to be able to find mappings from an input to an output from a larger non-linear function space [2]. This ability to represent a larger space of nonlinear

functions has shown to be very effective recently in many application domains such as natural language processing, computer vision, genomics, computer games and health[2]. Neural networks come in many flavors learning nonlinear mapping of different types of data such as Convolutional NNs being most effective with images and Recurrent NNs for time series and language data. Identifying the most effective neural network structure is one of the difficult and the most time-consuming aspect of applying neural networks to new application domains and data. Generally, neural networks try to exploit the relationships in the raw unstructured data (eg: image and text) presented to the network but with more structured data such as health records and ICD codes learning relationships is much complex. Our neural network models are mainly based on densely connected hidden layers with ReLu[6] activation function. We used the cross-entropy loss for the binary classification Adam optimizer. We used a simple feed forward network using Sklearn MLP classifier with hyperparameter tuning for the NN.

**Boosted Learning Algorithms**

Boosting is a process to ensemble multiple base learning algorithms to arrive at better overall performance than any individual base learner[1]. In contrast to independently building multiple models from the subsets of the data, boosting re-weights the training data every time a model is learned for future models. This weighting happens to give more preference to currently misclassified data points in the next round compared to the correctly classified data points. Therefore future learners try to do better on the misclassified data points leading to a collection base learners having a better-combined prediction. This process is sequential so each base learner is dependent on the output of the previously trained model (it is worthy to note XGBoost provides a parallel tree boosting alternative). In our work, we have experimented with several boosting meta-learning algorithms such as XGBoost[7], AdaBoost[5] and GBM[5]. XGBoost uses a variant of trees as the base learner whereas AdaBoost (from Sci-kit learn) can use many ML algorithms as base learners. GBM uses logistic regression by default as the base learner. We used all 3 types of boosting with tuned hyperparameters for comparison.

**Naive Bayes**

Naive Bayes is based on the Bayes theorem with a strong independence assumption between the covariates[1]. This assumption helps in building a simple probabilistic model for learning and inference. Naive Bayes coefficients scale linearly with the number of covariates making this a suitable model for high-dimensional data. We used Naive Bayes as a simple baseline learning algorithm for comparison.

**Support Vector Machines (SVM)**

SVMs[4] are maximum margin classifiers optimizing for learning a hyperplane having the maximum distance away from each of the class data points[1]. SVM is a linear classifier but with the kernel trick to map the inputs to the higher dimensional space, it can learn nonlinear decision boundaries in the input space. SVMs are very effective binary classifiers with the kernel trick[1]. With larger datasets, SVMs tend to become more computationally intensive.

eTable 1. Diagnostic codes used to exclude patients who had cancer, were pregnant, or were under palliative care.

| **Condition** | **ICD 9** | **ICD 10** |
|---|---|---|
| **Cancer** | 140.x - 239.x | C00.x - C99.x, D00.x - D49.x |
| **Pregnancy** | 630.x - 679.x | O00.x - O99.x |
| **Palliative** | V66 | Z51.0, Z51.1, Z51.5 |

eTable 2. Diagnostic codes used to identify the defined study outcome from emergency visit, hospitalization and death data.

| **ICD 10** | **Condition** |
|---|---|
| **T40.x** | Poisoning by, adverse effect of and underdosing of narcotics and psychodysleptics |
| **F55.x** | Abuse of non-psychoactive substances |
| **F11.x - F19.x** | Mental and behavioral disorders due to psychoactive substance use |

eTable 3. Baseline characteristics of study patients (n=392,979). Co-morbidities were determined using Elixhauser criteria. All p-values in the chi$^2$ test of independence were <0.001 unless otherwise indicated.

| Characteristic | Number without Event n=386,371 | Percent | Number with Event n=6,608 | Percent |
|---|---|---|---|---|
| **Age:** | | | | |
| Mean (SD) | 48.1 (16.4) | -- | 41.2 (12.4) | -- |
| 18-45 | 162057 | 41.9 | 3466 | 52.4 |
| 45-65 | 154632 | 40.0 | 2656 | 40.2 |
| >65* | 69682 | 18.0 | 486 | 7.4 |
| **Male** | 197491 | 50.3 | 3922 | 59.4 |
| **Female** | 194794 | 49.7 | 2686 | 40.6 |
| **Alcohol Disorder** | 66320 | 16.9 | 5220 | 79.0 |
| **Arrhythmia** | 90621 | 23.1 | 1959 | 29.6 |
| **Blood Loss Anemia** | 1164 | 0.3 | 82 | 1.2 |
| **Congestive Heart Failure** | 18954 | 4.8 | 565 | 8.6 |
| **Coagulopathy** | 8053 | 2.1 | 356 | 5.4 |
| **Deficiency Anemia** | 34188 | 8.7 | 971 | 14.7 |
| **Depression** | 159140 | 40.6 | 5518 | 83.5 |
| **Diabetes**** | 64132 | 16.3 | 1408 | 21.3 |
| **Substance Abuse Disorder** | 74678 | 19.0 | 5485 | 83.0 |
| **Fluid Disorder** | 42690 | 10.9 | 3012 | 45.6 |
| **Hypertension**** | 140171 | 35.7 | 2624 | 39.7 |
| **Hypothyroidism** | 45519 | 11.6 | 601 | 9.1 |
| **Injury^** | 195688 | 49.9 | 5541 | 83.9 |
| **Liver Disorder** | 21656 | 5.5 | 1588 | 24.0 |
| **Neurologic Disorder** | 230490 | 58.8 | 5387 | 81.5 |
| **Obesity** | 63393 | 16.2 | 970 | 14.7 |
| **Poisoning^** | 17434 | 4.4 | 2775 | 42.0 |
| **Psychoses** | 35870 | 9.1 | 3162 | 47.9 |
| **Renal Disorder** | 16166 | 4.1 | 499 | 7.6 |
| **Rheumatoid Conditions** | 111458 | 28.4 | 3157 | 47.8 |
| **HIV Infection** | 1098 | 0.3 | 141 | 2.1 |
| **Paralysis** | 3874 | 1.0 | 187 | 2.8 |
| **Peptic Ulcer Disease** | 11728 | 3.0 | 509 | 7.7 |
| **Pulmonary Circulation Disorder** | 9611 | 2.4 | 430 | 6.5 |
| **Chronic Pulmonary Disease** | 102990 | 26.3 | 2913 | 44.1 |
| **Peripheral Vascular Disease** | 14467 | 3.7 | 389 | 5.9 |
| **Valvular Disease** | 7308 | 1.9 | 226 | 3.4 |
| **Weight Loss** | 16207 | 4.1 | 747 | 11.3 |

*p-value for age >65 is an estimated 0.037

^ Injury: ICD10: S00-T98; Poisoning: ICD10: T36-T50
** Complicated, uncomplicated diabetes and hypertension were collapsed into one category each

eTable 4. Characteristics of study participants between training and validation groups using 2017 data.

| Characteristic | Number in training group N=275,150~ | Percent | Number in validation group N=117,829~ | Percent |
|---|---|---|---|---|
| **Age:** | | | | |
| **Mean (SD)** | 48.3 (16) | -- | 48.2 (16) | -- |
| **18-45** | 114356 | 41.5 | 49909 | 42.3 |
| **45-65** | 111859 | 40.7 | 47132 | 40.0 |
| **>65** | 48935 | 17.8 | 20788 | 17.6 |
| **Male** | 138603 | 48.5 | 59339 | 48.4 |
| **Female** | 136545 | 47.8 | 58490 | 47.7 |
| **Alcohol Disorder** | 46792 | 16.4 | 20199 | 16.5 |
| **Arrhythmia** | 63637 | 22.3 | 27201 | 22.2 |
| **Blood Loss Anemia** | 839 | 0.3 | 336 | 0.3 |
| **Congestive Heart Failure** | 13320 | 4.7 | 5694 | 4.6 |
| **Coagulopathy** | 5697 | 2.0 | 2393 | 2.0 |
| **Deficiency Anemia** | 24096 | 8.4 | 10179 | 8.3 |
| **Depression** | 112080 | 39.2 | 47628 | 38.9 |
| **Diabetes**** | 45131 | 15.8 | 19144 | 15.6 |
| **Substance Abuse Disorder** | 52609 | 18.4 | 22713 | 18.5 |
| **Fluid Disorder** | 30272 | 10.6 | 12780 | 10.4 |
| **Hypertension**** | 98546 | 34.5 | 41840 | 34.1 |
| **Hypothyroidism** | 31908 | 11.2 | 13666 | 11.2 |
| **Injury*** | 137423 | 48.1 | 58865 | 48.0 |
| **Liver Disorder** | 15252 | 5.3 | 6567 | 5.4 |
| **Neurologic Disorder** | 161706 | 56.5 | 69341 | 56.6 |
| **Obesity** | 44607 | 15.6 | 18882 | 15.4 |
| **Poisoning*** | 12503 | 4.4 | 5293 | 4.3 |
| **Psychoses** | 25422 | 8.9 | 10860 | 8.9 |
| **Renal Disorder** | 11403 | 4.0 | 4817 | 3.9 |
| **Rheumatoid Conditions** | 78268 | 27.4 | 33420 | 27.3 |
| **HIV Infection** | 774 | 0.3 | 336 | 0.3 |
| **Paralysis** | 2717 | 1.0 | 1176 | 1.0 |
| **Peptic Ulcer Disease** | 8239 | 2.9 | 3533 | 2.9 |
| **Pulmonary Circulation Disorder** | 6771 | 2.4 | 2877 | 2.3 |
| **Chronic Pulmonary Disease** | 72265 | 25.3 | 30949 | 25.3 |

| | | | | |
|---|---|---|---|---|
| **Peripheral Vascular Disease** | 10228 | 3.6 | 4278 | 3.5 |
| **Valvular Disease** | 5111 | 1.8 | 2215 | 1.8 |
| **Weight Loss** | 11477 | 4.0 | 4790 | 3.9 |

Note: p-values for chi$^2$ test of independence were all >0.06 when comparing training and validation sets.
*Injury: ICD10: S00-T98; Poisoning: ICD10: T36-T50
** Complicated, uncomplicated diabetes and hypertension were collapsed into one category each

eTable 5. Anatomical Therapeutic Chemical classification of opioid molecules used for this study and candidate predictors used to train ML algorithms.

| **Category (data source)** | **Description** |
|---|---|
| ATC codes used to identify opioids from PIN data | N01AH01, N01AH03, N01AH06, N07BC01, N07BC02, N07BC51, R05DA03, R05DA04, R05DA09, R05DA20, N02A |
| Opioid molecules used in this study | alfentanil, butorphanol, codeine, diamorphine, fentanyl, hydrocodone, hydromorphone, meperidine, morphine, oxycodone, oxymorphone, pentazocine, sufentanil, tapentadol, tramadol |
| Demographic information (PIN) | age, sex, postal codes, mean income |
| Drug utilization history (PIN) | drug dispenses in past 30 days using on ATC codes, oral morphine equivalents, concurrent use with benzodiazepines defined as at least 7 days of cumulative concurrent use in the 30 days prior to dispensation, number of dispensations and unique molecules of opioids and benzodiazepines |
| Health care utilization (PIN DAD) | flags for previous hospitalizations and emergency department visits, number of unique providers |
| ICD based co-morbidities (DAD, NACRS, Claims) | Elixhauser condition flags based on the past 5 years of claims, hospitalizations, and emergency visits. |

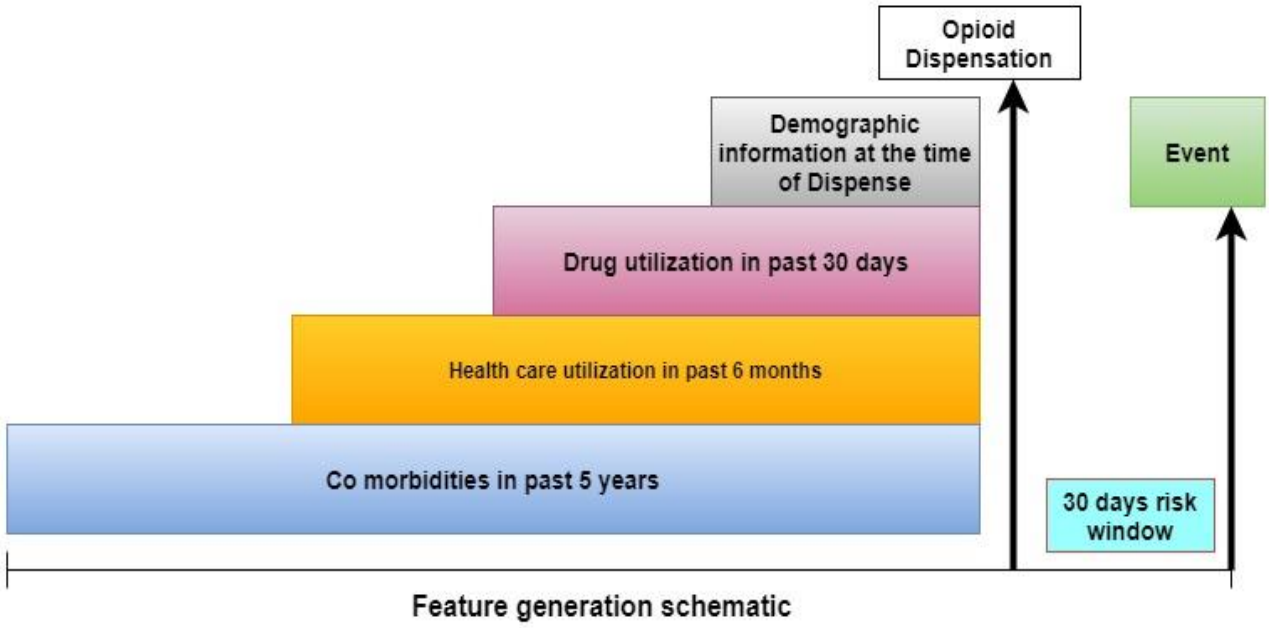
Note: ATC- Anatomical Therapeutic Chemical classification (https://www.whocc.no/atc_ddd_index); PIN- Pharmaceutical Information Network; ICD- International Statistical Classification of Diseases and Related Health Problems, World Health Organization; total number of features 283

eTable 6. Discrimination performance using area under the receiver operating characteristic curve (AUROC) of various ML algorithms using all features (demographics, health utilization, prescription history, co-morbidities). Training and validation were done using 2017 data (n=393,979); another independent validation was performed using 2018 data (n=393,023).

| Algorithm | Train | Validation 2017 | Validation 2018 |
|---|---|---|---|
| XGBoost Classifier | 0.897 | 0.870 | 0.884 |
| Logistic Regression | 0.887 | 0.869 | 0.884 |
| Gradient Boosting Classifier | 0.898 | 0.868 | 0.883 |
| AdaBoost Classifier | 0.884 | 0.868 | 0.882 |
| Random Forest Classifier | 0.909 | 0.863 | 0.881 |
| Ridge Classifier | 0.895 | 0.863 | 0.879 |
| SVM | 0.896 | 0.860 | 0.878 |
| Gaussian Naive Bayes | 0.846 | 0.826 | 0.847 |
| Decision Tree Classifier | 0.919 | 0.791 | 0.822 |
| Neural Networks | 0.827 | 0.804 | 0.821 |

Note: Logistic regression used L1 (lasso) parameter regularization

**eFigure 1. Schematic of study design and feature generation**



**eFigure2. Feature importance from logistic regression and tree-based XGBoost classifiers using the 2018 validation set. The logistic regression classifier relied more on co-morbidity data from DAD, NACRS, and Claims databases; XGBoost classifier relied more on data from the PIN database. AUROCs for both classifiers were similar at 0.88.**

| Logistic Regression | | XGBoost | |
|---|---|---|---|
| history of drug abuse | 1.00 | age at dispensation | 1.00 |
| age at dispensation | 0.65 | number of prescriptions dispensed in previous 30 days | 1.00 |
| history of prior hospitalization/ED visit | 0.62 | number of opioid dispensations in previous 30 days | 0.86 |
| history of alcohol use disorder | 0.62 | number of BZD dispensations in previous 30 days | 0.46 |
| history of fluid and electrolyte disorder | 0.32 | Doctor risk score* | 0.45 |
| history of poisoning | 0.31 | total OME consumed in previous 30 days | 0.43 |
| history of psychoses | 0.31 | history of poisoning | 0.37 |
| number of unique BZD dispensed in previous 30 days | 0.26 | pharmacy risk score** | 0.35 |
| history of depression | 0.19 | number of unique providers that prescribed an opioid or BZD | 0.34 |
| concurrent use of opioid and BZD in previous 30 days | 0.19 | income | 0.34 |
| history of injury | 0.17 | history of prior hospitalization/ED visit | 0.26 |

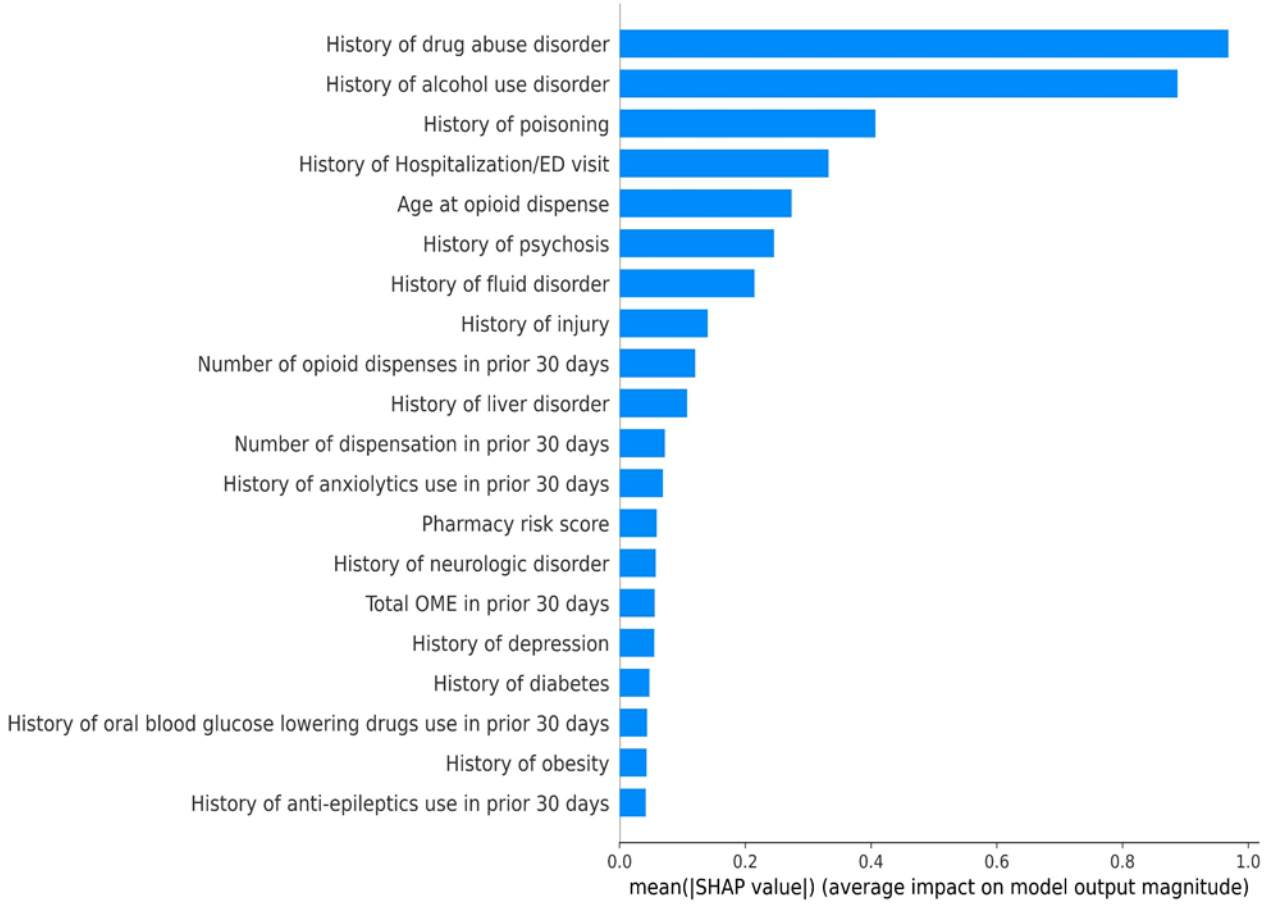
Note: Logistic regression used L1 (lasso) parameter regularization; BZD- benzodiazepine; OME- oral morphine equivalents; ED: emergency department

*derived feature using proportion of opioid/benzodiazepine patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each physician;

**derived feature using proportion of opioid/benzodiazepine patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each pharmacy

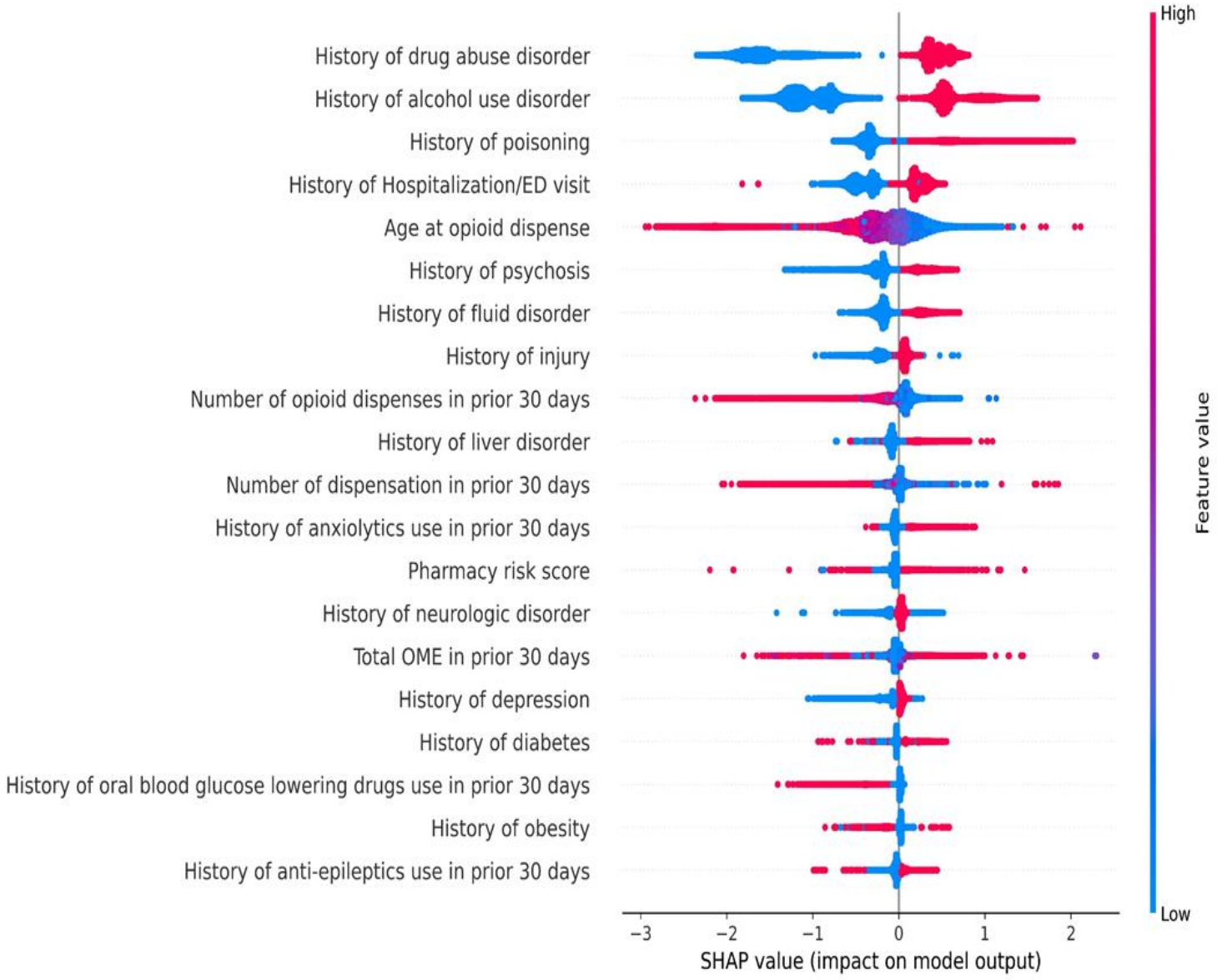
**eFigure 3. SHAP values and feature impact of the XGBoost classifier using the 2018 validation set to describe "associations" between features and the outcome. Features with the most impact on the model with drug abuse ranked highest (A); tornado plot illustrating feature impact (B); explaining the prediction of study outcome based on predictor values for 4 patients using SHAP values(C).**

**(A)**




Note: Pharmacy risk score- derived feature using proportion of opioid patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each pharmacy; training and validating the XGBoost classifier with these features alone resulted in an AUC of 0.877 in the 2018 validation set

**(B)**




Note: Pharmacy risk score- derived feature using proportion of opioid/benzodiazepine patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each pharmacy; red indicates higher values of categorical variables and plots to the right of 0.0 indicate the tendency to be associated with the study outcome while blue indicates lower values of categorical variables and plots to the left of 0.0 indicate the tendency to be associated with no outcome

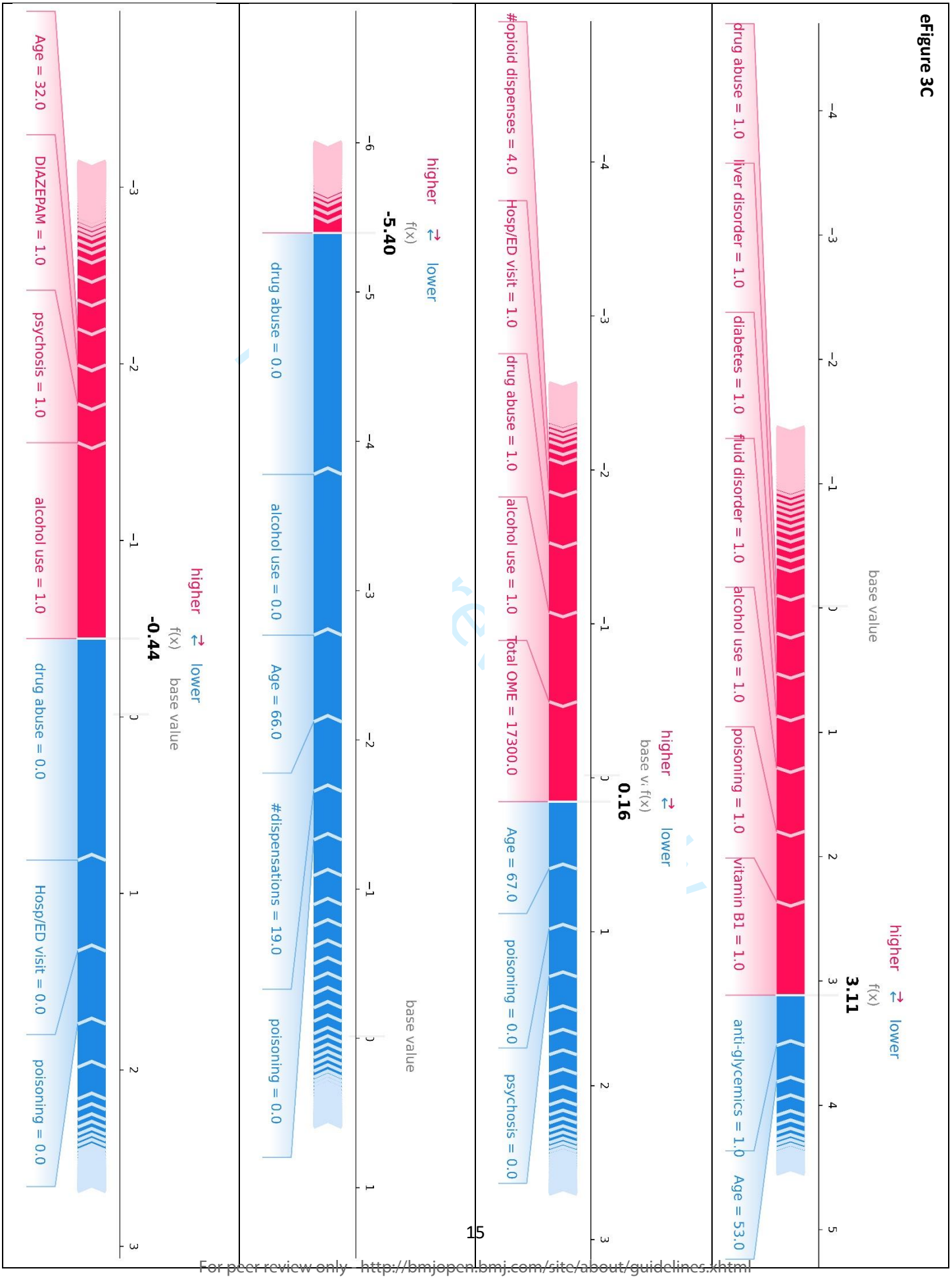
**(C)**

**How to read the figure on the next page:** Using hospitalization within 30-days of an opioid dispensation as the outcome of interest, there are 4 scenarios to consider: the XGBoost classifier has low or high confidence in predicting a hospitalization and low or high confidence in predicting **NO** hospitalization. Start at the base SHAP value of near 0.0 ("base value") in which the classifier is not confident in the prediction. SHAP values (in bold) that are above 0.0 indicate a tendency towards a hospitalization while those that are below 0.0 indicate a tendency for **NO** hospitalization. As the SHAP value moves above 0.0, for example 3.11 in the top panel, the classifier's confidence in predicting a hospitalization is higher. As the SHAP value approaches closer to the base value, for example 0.16 in the second panel, the classifier has relatively lower confidence in predicting a hospitalization. When the SHAP value is below 0.0, for example -5.4 in the third panel, the classifier's confidence in predicting **NO** hospitalization is higher and when the SHAP value is closer to 0.0, for example -0.44 in the bottom panel, the classifier has lower confidence in predicting **NO** hospitalization.

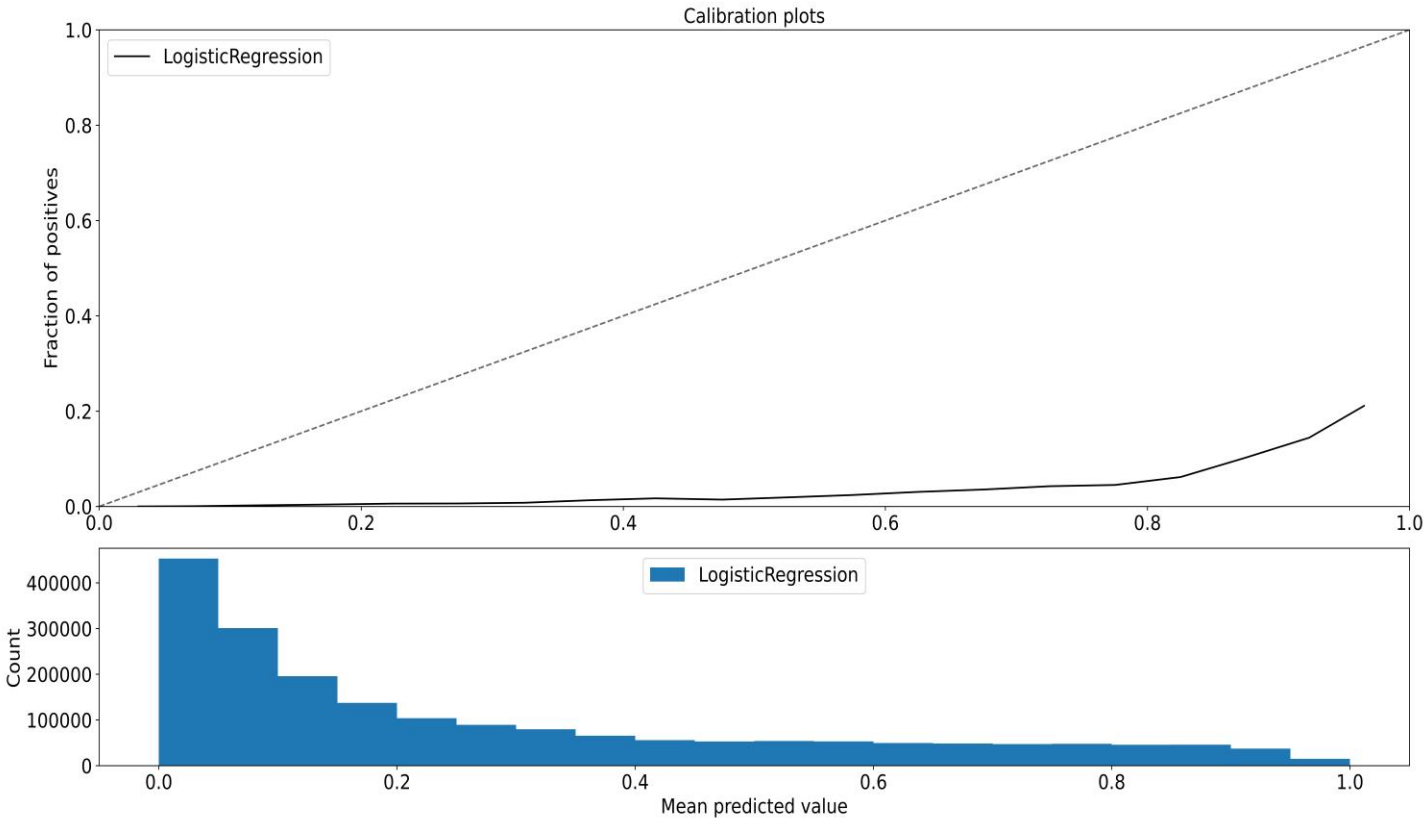
The top panel (SHAP value 3.11) depicts an instance predicted to be high risk for our outcome. This individual has a positive history of drug abuse disorder, liver disorder, diabetes, fluid/electrolyte disorder, alcohol use disorder, poisoning and B vitamin use in the prior 30 days. The third panel (SHAP value -5.40) depicts an instance predicted to be low risk (i.e., no hospitalization) and has a negative history for poisoning, drug and alcohol use disorder.

**Note-** drug abuse: drug abuse disorder; poisoning: history of poisoning; vitamin B1: vitamin B1 in prior 30 days; anti-glycemics: anti-glycemic agents in prior 30 days; age: age at opioid dispensation; # opioid dispenses: number of opioid dispensations in prior 30 days; Hosp/ED visit: history of prior hospitalizations and/or emergency visits in past 6 months; Total OME: total oral morphine equivalents in prior 30 days; DIAZEPAM: history of diazepam use in prior 30 days.

**eFigure 3C**



**eFigure 4.** Calibration curve plotting observed vs. quantiles of estimated risk for the logistic regression (L1) classifier using the 2018 validation dataset. The majority of counts (dispensations) were predicted to be lower risk.



**eReferences**

1. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning, vol. 1. Springer series in statistics New York (2001)
2. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
3. Zhu, H. Zou, S. Rosset, T. Hastie, “Multi-class AdaBoost”, 2009.
4. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2011 May 6;2(3):1-27.
5. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
6. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. InProceedings of the 27th international conference on machine learning (ICML-10) 2010 (pp. 807-814).
7. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).