

BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open

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

Journal:	<i>BMJ Open</i>
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 Kulkarni, Vinaykumar; OKAKI Health Analytics Eurich, Dean; University of Alberta, School of Public Health Kumar, Luke; Alberta Machine Intelligence Institute Samanani, Salim; Okaki Health Intelligence,
Keywords:	PUBLIC HEALTH, EPIDEMIOLOGY, Adverse events < THERAPEUTICS, Health & safety < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Clinical governance < HEALTH SERVICES ADMINISTRATION & MANAGEMENT

SCHOLARONE™
Manuscripts



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.

1
2
3 **Safe opioid prescribing: a machine learning approach to predicting 30-day risk after an opioid**
4 **dispensation in Alberta, Canada**
5

6 **Author list (in order):**
7

8 Vishal Sharma (0000-0001-7907-1183), Vinaykumar Kulkarni, Dean T. Eurich (0000-0003-2197-
9 0463), Luke Kumar, Salim Samanani (0000-0001-6751-4805)
10
11

12
13
14 **Address for each author:**
15

16 2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of
17 Alberta, Edmonton, Alberta, Canada, T6G 2E1 [Vishal Sharma BPharm PhD Candidate](#),
18
19

20
21 OKAKI Health Intelligence, Edmonton, Alberta, Canada, Vinaykumar Kulkarni MSc
22
23

24
25 2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of
26 Alberta, Edmonton, Alberta, Canada, T6G 2E1 Dean Eurich professor
27
28

29
30 Alberta Machine Intelligence Institute, Edmonton, Alberta, Canada, T5J 3B1 Luke Kumar MSc
31
32

33
34 OKAKI Health Intelligence, Calgary, Alberta, Canada, Salim Samanani MD, Medical Director
35
36

37
38 **Corresponding Author:**
39

40 Dean Eurich, 2-040 Li Ka Shing Center for Health Research Innovation, University of Alberta,
41 Edmonton, Alberta, Canada, T6G 2E1; Phone 780-492-6333; fax 780-492-7455; email:
42 deurich@ualberta.ca
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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.

Copyright/license for publication: *The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide licence to the Publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) licence any third party to do any or all of the above.*

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).

1
2
3 **Data Sharing:** The data used in this study is not available for external analysis. However, administrative
4 health data can be accessed from Alberta Health by following defined research protocols and
5 confidentiality agreements.
6
7

8
9 **Transparency:** The lead author, VS, (the manuscript's guarantor, Dean Eurich) affirms that the
10 manuscript is an honest, accurate, and transparent account of the study being reported; that no
11 important aspects of the study have been omitted; and that any discrepancies from the study as
12 originally planned (and, if relevant, registered) have been explained.
13
14

15
16 **Word Count: 2659**
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

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 crisis¹; a major part of the policy response to the opioid crisis focuses on endorsing safe, appropriate opioid prescribing²⁻⁴. 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 ICD⁸ 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 populations¹¹. 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.

1
2
3 Supervised machine learning (ML)^{12,13} is an approach that uses computer algorithms to
4
5 build predictive models in the clinical setting that can make use of the large amounts of
6
7 available administrative data^{14,15}, all within a well-defined process¹⁶. Supervised ML trains on
8
9 labelled data to develop prediction models that are specific to different populations and, in
10
11 many cases, can provide better predictive performance than traditional, population-based
12
13 statistical models^{10,15,17}. We identified one study¹⁰ that applied ML techniques to predict
14
15 overdose risk in opioid patients pursuant to a prescription. In their validation sample, they
16
17 found that the DNN (deep neural network) and GBM (gradient boosting machine) algorithms
18
19 carried the best discrimination performance based on estimated c-statistics and that the ML
20
21 approach out-performed the guideline approach in terms of predictive performance.
22
23
24
25
26
27

28 The objective of our study was to develop and validate ML algorithms to predict the 30-
29
30 day risk of hospitalization, emergency visit and mortality for a patient in Alberta, Canada at the
31
32 time of an opioid dispensation using administrative data routinely available to health
33
34 departments and PMPs. We hypothesized that the ML process would perform better than the
35
36 current guideline approach for predicting risk of adverse outcomes related to opioid
37
38 prescribing.
39
40
41
42
43

44 **Methods**

47 **Study Design and Participants**

48
49
50 This prognostic study used a supervised ML scheme. All patients in Alberta, Canada who
51
52 received a dispensation for an opioid, were 18 years of age and older between Jan 1, 2017 and
53
54 Dec 31, 2018 were eligible. Patients were excluded if they had any previous diagnosis of
55
56
57
58
59
60

1
2
3 cancer, received palliative interventions or were pregnant during the study period (eTable 1 in
4 Supplement) as use of opioids in these contexts is clinically different.
5
6

7
8 Government health departments and payers in many jurisdictions have systems to capture
9 prescription histories and ICD diagnostic codes. As such, we linked various administrative
10 health data sets available in Alberta, Canada using unique patient identifiers in order to
11 establish a complete description of patient demographics, drug exposures and health
12 outcomes. These databases include 1) *Pharmaceutical Information Network (PIN)*: PIN data
13 includes all dispensing records from community pharmacies from all prescriber types occurring
14 in the province outside of the hospital setting. PIN collects all drug dispensations irrespective of
15 age or insurance status in Alberta, 2) *Population and Vital Statistics Data (VS, Alberta Services)*:
16 sex, age, date of birth, death date, immigration and emigration data, and underlying cause of
17 death according to the World Health Organization algorithm using ICD codes⁸, 3)
18 *Hospitalizations and Emergency Department Visits (NACRS [National Ambulatory Care*
19 *Reporting System], DAD [Discharge Abstract Database])*: all services, length of stay, diagnosis
20 (up to 25 ICD-10⁸ based diagnoses). Data and coding accuracy are routinely validated both
21 provincially and centrally via the Canadian Institute for Health Information, and 4) *Physician*
22 *Visits/Claims (Alberta Health)*: date of service, ICD code associated with the claim, procedure
23 and billing information.
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

47
48 This study followed the TRIPOD and STARD reporting guidelines¹⁸⁻²⁰ and received ethics
49 approval from the University of Alberta ethics board (Pro00083807_AME2). All analyses were
50 done using Python, version 3.7 (Python Software Foundation).
51
52
53
54
55
56
57

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 codes²¹), co-morbidity history using ICD-based Elixhauser score categories²², health care utilization (number of unique opioid prescribers, number of hospital visits), and drug utilization (level 3 ATC codes²³, oral morphine equivalents²⁴, 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%) sets¹³. Baseline characteristics and event rates were compared in the training vs validation group, and between those who experienced the outcome and those who

1
2
3 did not, using chi-squared tests of independence. As well, we used all 2018 data as another
4
5 independent validation set.
6
7

8
9 First, we trained commonly used¹³ ML algorithms (eAppendix in Supplement) and tuned
10
11 model hyperparameters using k-fold (k=5) cross validation to address model overfitting^{13,25}. As
12
13 is common in ML validation studies¹⁰, we reported model discrimination performance using
14
15 area under the receiver operating characteristic curve (AUROC; c-statistic), positive predictive
16
17 value (PPV), positive likelihood ratios (PLR), number needed to screen (NNS) and plotted
18
19 AUROC and precision-recall curves (PRCs). For the more interpretable XGBoost and logistic
20
21 regression classifiers, we reported feature importance²⁶ and plotted PRCs that compared all
22
23 dispenses to those within the top 10 percentiles of estimated risk. As well, for the XGBoost
24
25 classifier, we described feature impact on model outcome using SHAP values^{27,28} to add an
26
27 additional layer of interpretability. Calibration is crucial in the process of developing a risk
28
29 predictor²⁹ so we assessed calibration performance on the 2018 data by dividing the study
30
31 cohort into percentile categories according to the predicted risk of a dispensation, as was done
32
33 in previous studies^{10,30}. Using the XGBoost and logistic regression classifiers, we analyzed the
34
35 top 0.1, 1, 5, and 10 percentiles of predicted risk by the number of true and false positives,
36
37 positive likelihood ratios, post-test probabilities, and number needed to screen. We also
38
39 performed a simulation of daily data uploads for 2018 Quarter 1 to view the predictive power if
40
41 a ML risk predictor were to be deployed into a monitoring workflow.
42
43
44
45
46
47
48
49

50
51 We then compared ML risk prediction to current guideline approaches as others have¹⁰,
52
53 using the 2019 Centers for Medicare & Medicaid Services opioid safety measures³¹ and the
54
55 2017 Canadian Opioid Prescribing Guideline³. As well, we compared the discrimination
56
57

1
2
3 performance of different logistic regression classifier models using various combinations of
4
5 features derived from their respective databases: **1)** demographic and drug/health utilization
6
7 features from PIN and **2)** co-morbidity features derived from DAD, NACRS and Claims.
8
9

10 11 **Patient and Public Involvement**

12
13
14 This research was done without patient involvement. Patients were not invited to comment on
15
16 the study design and were not consulted to develop patient relevant outcomes or interpret the
17
18 results. Patients were not invited to contribute to the writing or editing of this document for
19
20 readability or accuracy. There are no plans to disseminate the results of the research to study
21
22 participants.
23
24
25

26 27 **Results**

28 29 30 **Patient Characteristics and Predictors**

31
32
33 We identified 392,979 patients with at least one opioid dispensation in 2017 (Figure 1). This
34
35 cohort was used to train (n= 275,150, 70%) and validate (n=117,829, 30%) ML models. In 2017
36
37 and 2018, 6,608 and 5,423 patients experienced the defined outcome, respectively. Baseline
38
39 characteristics were different between those who experienced the outcome and those who did
40
41 not (eTable 3 in Supplement) while characteristics were similar between the training and
42
43 validation sets (eTable 4 in Supplement). There were 2,283,075 opioid dispensations in 2017
44
45 and 1,977,389 in 2018. Overall, in 2017, 2.03% (n= 45,757) of opioid dispensations were
46
47 associated with the outcome; in 2018, the estimate was 1.6% (n= 31,392).
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 As described above, we categorized our candidate features into four groups (eTable 5 in
4
5 Supplement).

6 7 8 **Machine-Learning Prediction Performance**

9
10
11 Using the 2017 validation set, AUROCs for the XGBoost and logistic regression classifiers had
12
13 the highest discrimination performance at 0.87, while the neural network classifier had lower
14
15 performance at 0.80 (eTable 6 in Supplement).

16
17
18
19
20 Discrimination performance was similar for the 2018 validation set (n=393,023; eTable 6
21
22 in Supplement). XGBoost and logistic regression had the highest estimated AUROCs while the
23
24 neural network classifier was lower (Figure 2A). As expected, precision-recall curves indicate
25
26 stronger predictive power in opioid dispensations at higher predicted risk percentiles (Figure
27
28 2C, 2D).

29
30
31
32
33 In the 2018 validation set, although discrimination performance was similar (0.88),
34
35 individual feature importance was different between the logistic regression and XGBoost
36
37 classifiers, with logistic regression feature importance more reliant on co-morbidity data from
38
39 DAD, NACRS and Claims while XGBoost relied more on drug utilization data from PIN (eFigure
40
41 2). In the XGBoost classifier, history of drug abuse, alcoholism, and prior hospitalization carried
42
43 the highest impact for predicting the study outcome (eFigure 3A) where the presence of these
44
45 features in a patient suggested a strong tendency towards having the defined outcome (eFigure
46
47 3B and 3C).

48 49 50 51 52 **Calibration**

1
2
3 When considering dispensations predicted to be in the highest percentiles of risk, the top 5-
4
5 percentile captured 42% of all outcomes using the XGBoost and logistic regression classifiers
6
7 (Table 1). Also, as the predicted risk percentiles get higher (top 10 percentile to top 0.1
8
9 percentile), so too do the corresponding PPVs with the top 0.1 percentile associated with a PPV
10
11 of 33% for the XGBoost classifier. As well, lower categories of risk percentiles were associated
12
13 with lower outcomes (Figure 3A and 3B). When we simulated a monitoring workflow scenario
14
15 with daily data uploads, a similar pattern was illustrated where the dispensations predicted to
16
17 be higher risk had higher event rates (Figure 3C and 3D).
18
19
20
21
22

23 After using the XGBoost and logistic regression classifiers to identify the dispensations in the
24
25 highest predicted risk percentiles, the pre-test probability of the outcome (1.6%) was
26
27 transformed into higher post-test probabilities, with higher probabilities in the riskier
28
29 percentiles (Table 1). The number needed to screen also decreased as predicted risk increased
30
31 (Table 1).
32
33
34
35

36 Comparing discrimination performance, ML risk prediction outperformed the current
37
38 guideline approaches when using various combinations of guideline recommendations (Table
39
40 2). In many of the guideline scenarios, the estimated AUROCs were close to the 0.5 mark.
41
42 When we estimated the discrimination performance of the logistic regression classifier based
43
44 on database source, using all databases produced an AUROC of 0.88. Reducing the database
45
46 source to only DAD, NACRS, Claims (co-morbidities only) resulted in an AUROC of 0.85, while
47
48 PIN (prescription history) only was 0.78 (Table 3).
49
50
51
52
53

54 Discussion

55
56
57
58
59
60

1
2
3 This study showed that ML techniques using available administrative data (prescription
4 histories and ICD codes) may provide enough discriminatory power to predict adverse
5 outcomes associated with opioid prescribing. Indeed, our ML analyses showed very high
6 discrimination performance at 0.88. The linear model (logistic regression) and XGBoosted Trees
7 carried higher discrimination and calibration performance, while the neural network classifier
8 did not perform as well. By identifying the predicted top 5-10 percentile of absolute risk
9 pursuant to an opioid dispensation, we were able to capture approximately half of all outcomes
10 using ML methods. All ML models we trained had higher discrimination performance using
11 independent (external) validation sets than the clinical guideline approach.
12
13
14
15
16
17
18
19
20
21
22
23
24
25

26 Since the prevalence of our defined outcome is relatively low in the general population,
27 PPVs would also be expectedly low. However, estimated PPVs increased when we considered
28 higher risk dispensations, as is expected since PPV is related to event prevalence. This is
29 important because different users of a risk predictor will require different predictive
30 capabilities. Similarly, our estimates of positive likelihood ratios and associated post-test
31 probabilities also increased in dispensations with higher predicted risk indicating the strong
32 predictive power of the XGBoost and logistic regression classifiers; likelihood ratios >10
33 generate conclusive changes from pre-test to post-test probabilities³².
34
35
36
37
38
39
40
41
42
43
44
45

46 The current guideline approach to assess absolute opioid prescribing risk produced c-
47 statistic estimates closer to 0.5 indicating that discrimination was not much better than chance
48 alone. ML models with higher predictive power can better support health departments and
49 PMPs with monitoring mandates to identify and intervene on those at high risk and their
50 associated prescribers. We also found that adding co-morbidity features from administrative
51
52
53
54
55
56
57

1
2
3 databases increased prediction performance compared to prescription history alone, thus
4
5 making the case for the use of this data by PMPs and health departments. However, if only
6
7 prescription history is available, our trained XGBoost classifier still had strong discrimination
8
9 performance.
10
11

12
13 We found only one study that used ML approaches to quantify the absolute risk of an
14
15 event pursuant to an opioid dispensation¹⁰. Their methodology used rolling 3-month windows
16
17 for estimating risk and ML model training while we used historic records to estimate 30-day
18
19 risk. Differences in study population and feature selection may explain why their highest
20
21 performing ML model was deep learning (neural network classifier) and ours was not.
22
23 Nevertheless, we were able to replicate and build upon their discrimination performance using
24
25 our ML approach as we both were able to show that ML approaches have higher predictive
26
27 power than guideline approaches. Both of our studies used predicted percentile risk estimates
28
29 to identify high risk dispensations and were able to do so with strong discrimination and
30
31 calibration performance. This is important because interventions can be targeted to higher risk
32
33 instead of lower risk patients. Another study we found describes how identifying cases in higher
34
35 predicted risk percentiles using ML methods can be deployed in hospital settings for the
36
37 purpose of targeted interventions³⁰ upon discharge.
38
39
40
41
42
43
44
45

46 The limitations of our study are similar to other ML studies¹⁰ and need to be addressed
47
48 when considering deployment of ML risk predictors. Our training dataset was not able to
49
50 account for non-prescription opioid consumption and the risk associated with non-prescription
51
52 use, both of which are substantial contributors to overall risk². Regarding our analysis, we
53
54 assumed that all dispensations were independent events; future research in this area should
55
56
57

1
2
3 focus on employing ML methods using correlated data. As with all ML projects, our models
4
5 were trained using Alberta data and might not be generalizable to other populations, or to
6
7 specific populations within Alberta. However, our analyses were done on a large population
8
9 and these results would be expected to be generalizable to the vast majority of patients.
10
11 Moreover, one of the benefits of the ML process is that models can be retrained or similar
12
13 methods could be used to develop new models to accommodate different populations.
14
15
16
17

18 This study suggests that ML risk prediction can support PMPs, especially if able to use
19
20 administrative health data. The ML process allows for model training, validation and
21
22 deployment to specific settings. However, uptake of this technology is limited for the time
23
24 being. Further research can assess whether implementation of a ML-based monitoring system
25
26 by PMPs leads to improved clinical outcomes.
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 Pain Opioid 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. Hsieh 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.

- 1
 - 2
 - 3
 - 4
 - 5
 - 6
 - 7
 - 8
 - 9
 - 10
 - 11
 - 12
 - 13
 - 14
 - 15
 - 16
 - 17
 - 18
 - 19
 - 20
 - 21
 - 22
 - 23
 - 24
 - 25
 - 26
 - 27
 - 28
 - 29
 - 30
 - 31
 - 32
 - 33
 - 34
 - 35
 - 36
 - 37
 - 38
 - 39
 - 40
 - 41
 - 42
 - 43
 - 44
 - 45
 - 46
 - 47
 - 48
 - 49
 - 50
 - 51
 - 52
 - 53
 - 54
 - 55
 - 56
 - 57
 - 58
 - 59
 - 60
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.whooc.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 systems 2017.
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 equivalents²⁴, 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

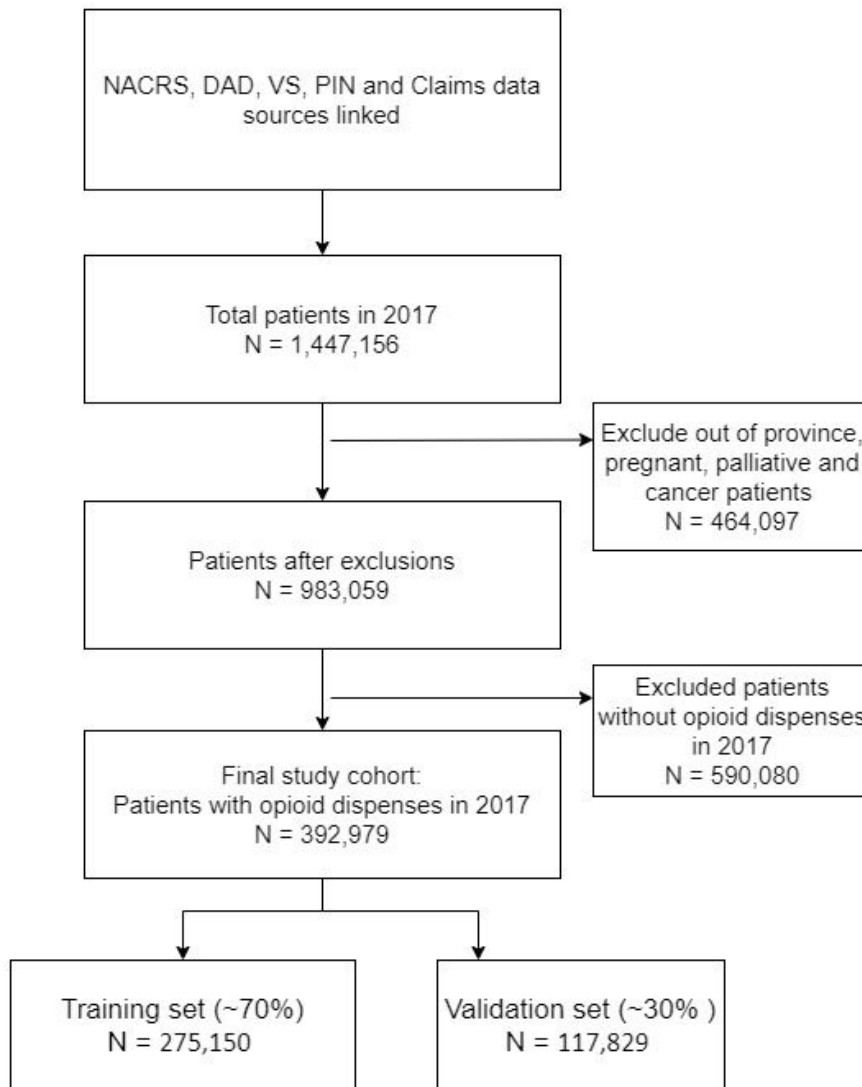
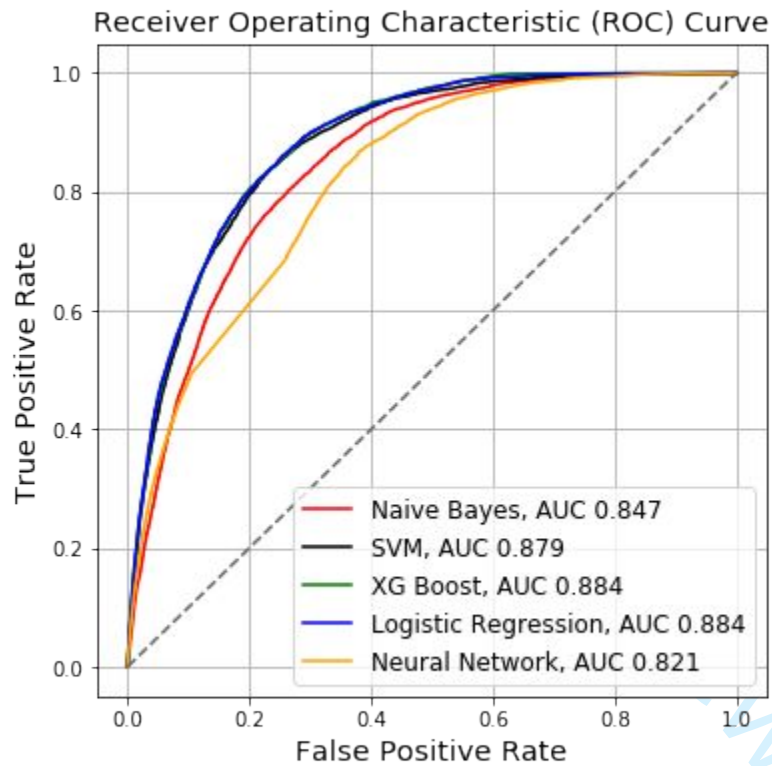
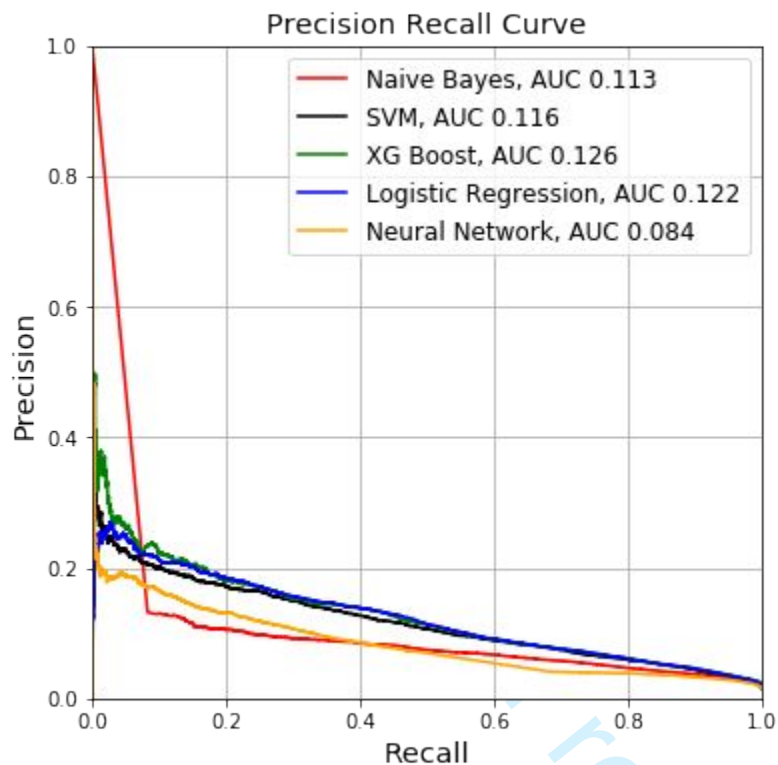


Figure 2. Area under the receiver operating characteristic curve (AUROC) (A) and precision-recall curves (B) for all dispensations using logistic regression, neural network, support vector machine (SVM), XGBoost and Naïve-Bayes; precision-recall curves for higher risk dispensations according to predicted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

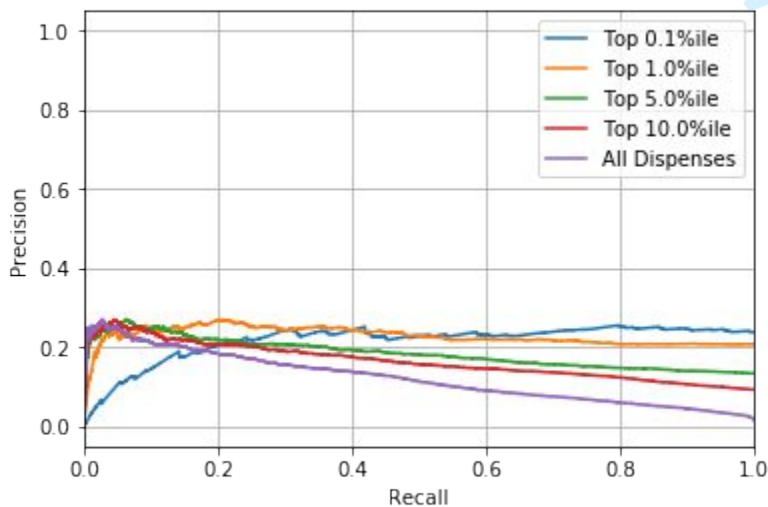
(A)



(B)

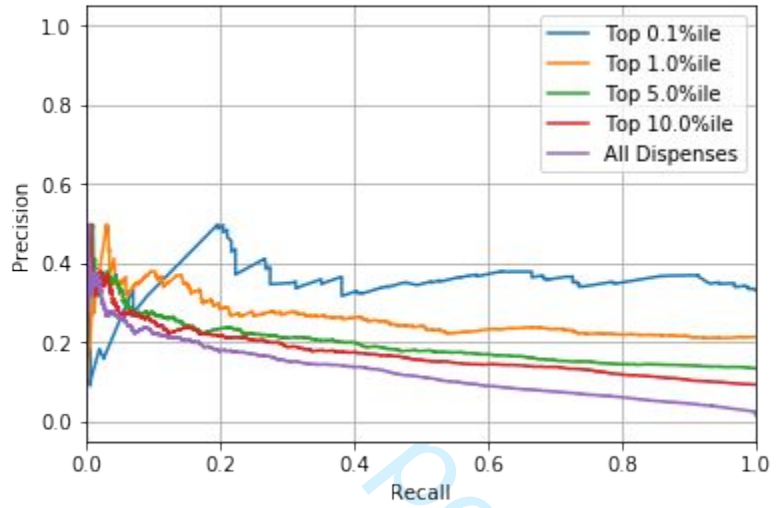


(C) Logistic Regression



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

(D) XGBoost

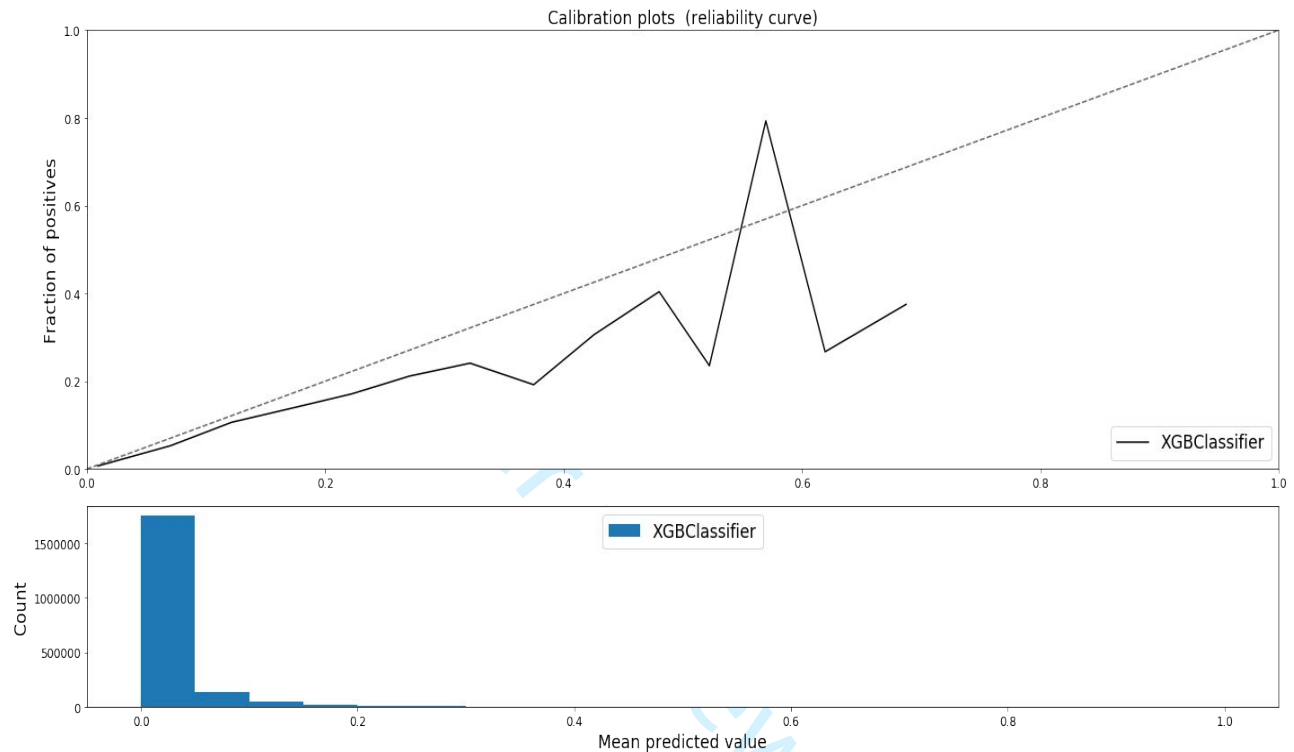


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

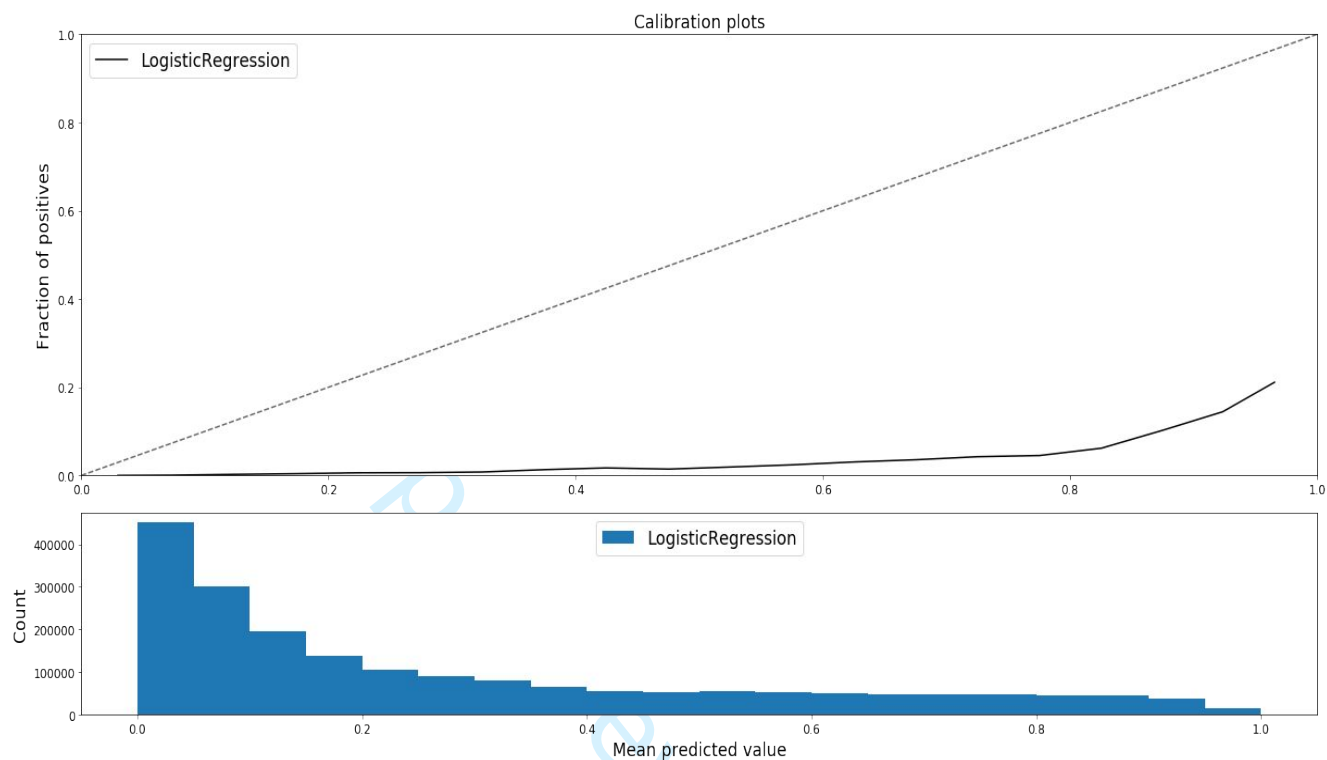
For peer review only

1
2
3 Figure 3. Calibration curves plotting: **1**) observed vs. quantiles of estimated risk for XGBoost (A) and
4 logistic regression (B) classifiers using the 2018 validation dataset and **2**) simulation of a clinical
5 workflow with daily uploads and events per 100 daily dispenses by risk percentiles using 2018 Quarter 1
6 (Q1) data for logistic regression (C) and XGBoost (D) classifiers. For both classifiers, the majority of
7 counts (dispensations) were predicted to be lower risk.
8
9

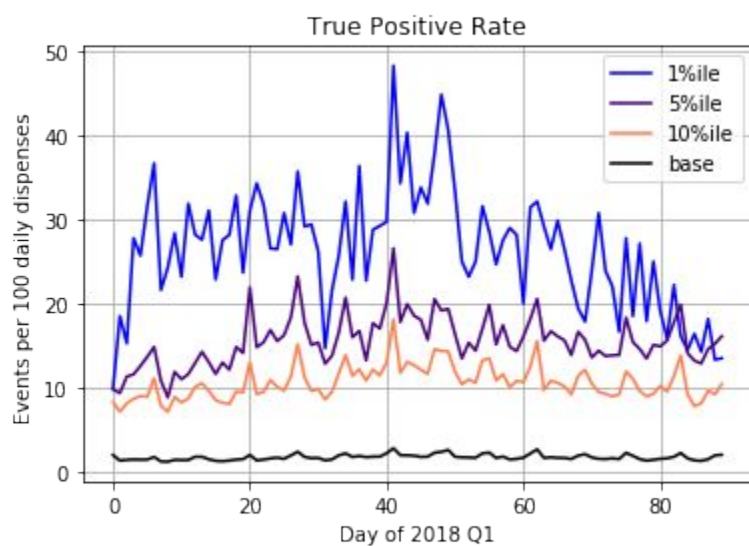
10 (A) XGBoost



(B) Logistic Regression

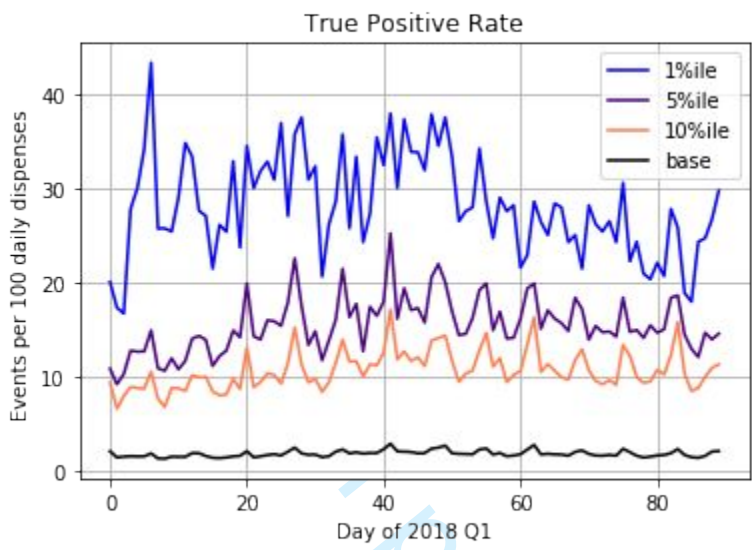


(C) Logistic Regression



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

(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 chi² 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

1
2
3 from a larger non-linear function space [2]. This ability to represent a larger space of nonlinear
4 functions has shown to be very effective recently in many application domains such as natural
5 language processing, computer vision, genomics, computer games and health[2]. Neural
6 networks come in many flavors learning nonlinear mapping of different types of data such as
7 Convolutional NNs being most effective with images and Recurrent NNs for time series and
8 language data. Identifying the most effective neural network structure is one of the difficult and
9 the most time-consuming aspect of applying neural networks to new application domains and
10 data. Generally, neural networks try to exploit the relationships in the raw unstructured data (eg:
11 image and text) presented to the network but with more structured data such as health records
12 and ICD codes learning relationships is much complex. Our neural network models are mainly
13 based on densely connected hidden layers with ReLu[6] activation function. We used the cross-
14 entropy loss for the binary classification Adam optimizer.
15
16
17
18

19 **Boosted Learning Algorithms**

20
21 Boosting is a process to ensemble multiple base learning algorithms to arrive at better overall
22 performance than any individual base learner[1]. In contrast to independently building multiple
23 models from the subsets of the data, boosting re-weights the training data every time a model is
24 learned for future models. This weighting happens to give more preference to currently
25 misclassified data points in the next round compared to the correctly classified data points.
26 Therefore future learners try to do better on the misclassified data points leading to a collection
27 base learners having a better-combined prediction. This process is sequential so each base
28 learner is dependent on the output of the previously trained model (it is worthy to note XGBoost
29 provides a parallel tree boosting alternative). In our work, we have experimented with several
30 boosting meta-learning algorithms such as XGBoost[7], AdaBoost[5] and GBM[5]. XGBoost uses
31 a variant of trees as the base learner whereas AdaBoost (from Sci-kit learn) can use many ML
32 algorithms as base learners. GBM uses logistic regression by default as the base learner. We used
33 all 3 types of boosting with tuned hyperparameters for comparison.
34
35
36
37
38

39 **Naive Bayes**

40
41 Naive Bayes is based on the Bayes theorem with a strong independence assumption between the
42 covariates[1]. This assumption helps in building a simple probabilistic model for learning and
43 inference. Naive Bayes coefficients scale linearly with the number of covariates making this a
44 suitable model for high-dimensional data. We used Naive Bayes as a simple baseline learning
45 algorithm for comparison.
46
47
48

49 **Support Vector Machines (SVM)**

50
51 SVMs[4] are maximum margin classifiers optimizing for learning a hyperplane having the
52 maximum distance away from each of the class data points[1]. SVM is a linear classifier but with
53 the kernel trick to map the inputs to the higher dimensional space, it can learn nonlinear decision
54 boundaries in the input space. SVMs are very effective binary classifiers with the kernel trick[1].
55 With larger datasets, SVMs tend to become more computationally intensive.
56
57

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For peer review only

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 χ^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.

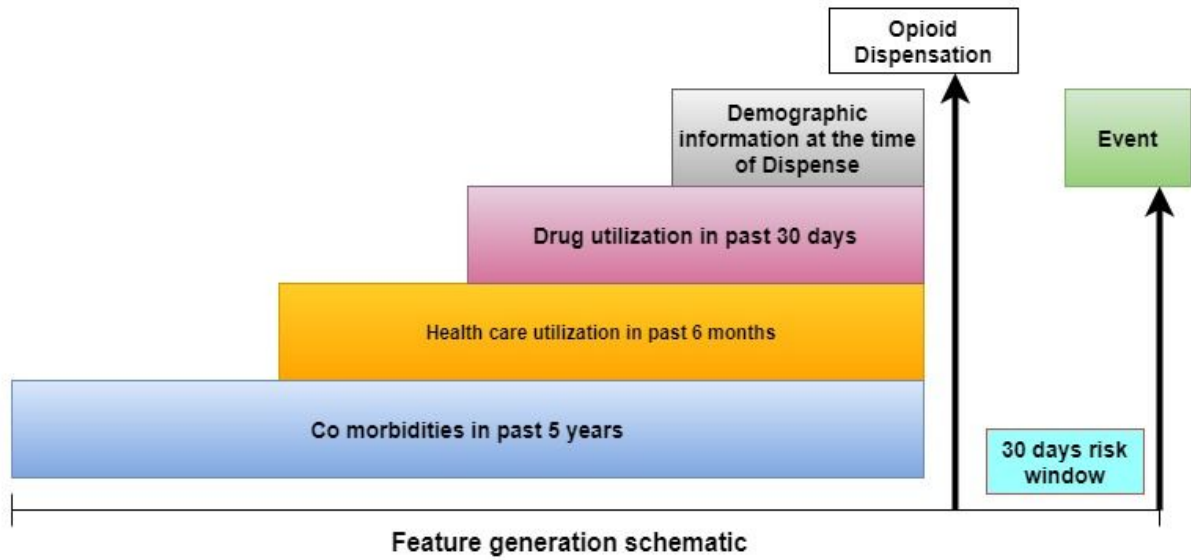
Category (data source)	Description
Demographic information (PIN)	age, sex, postal codes, mean income
Drug utilization history (PIN)	drug dispenses in past 30 days using on ATC codes, oral morphine equivalents, concurrent use with benzodiazepines, number of dispensations and unique molecules of opioids 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



review only

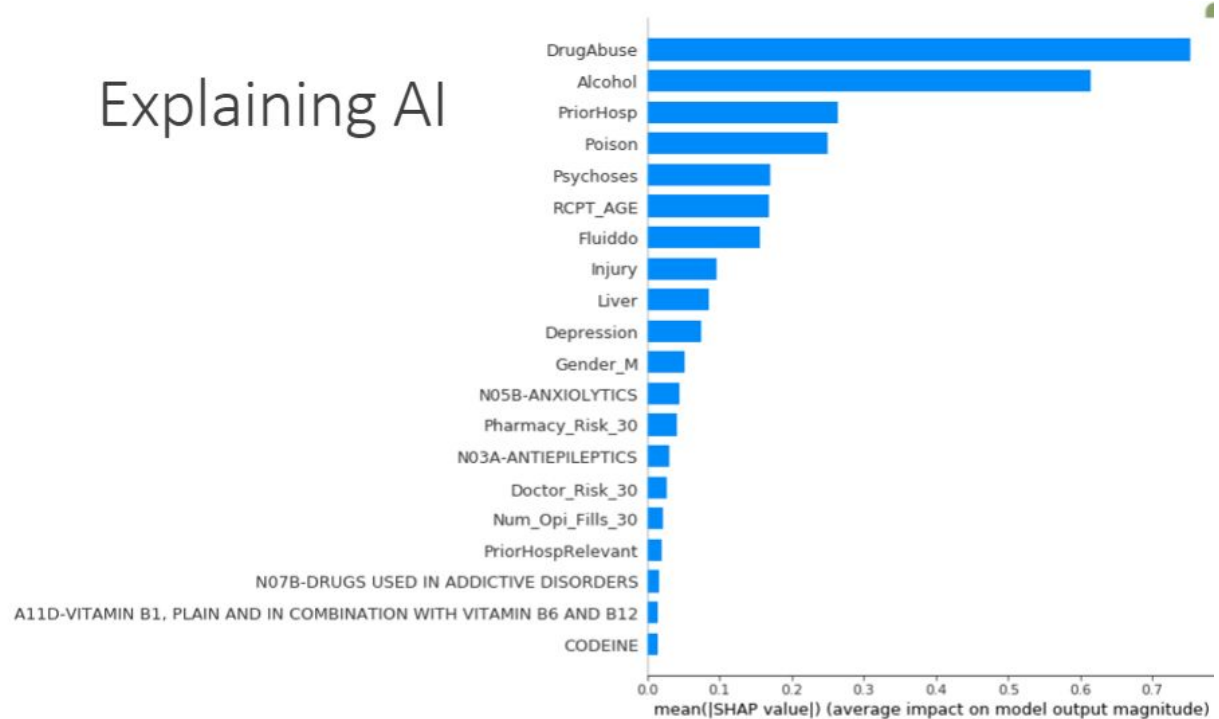
eFigure2. Feature importance from logistic regression and tree-based (XGBoost) classifiers using the 2018 validation set. The logistic regression classifier relied more on co-morbidity data from DAD, NACRS, and Claims databases; XGBoost classifier relied more on data from the PIN database. AUROCs for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

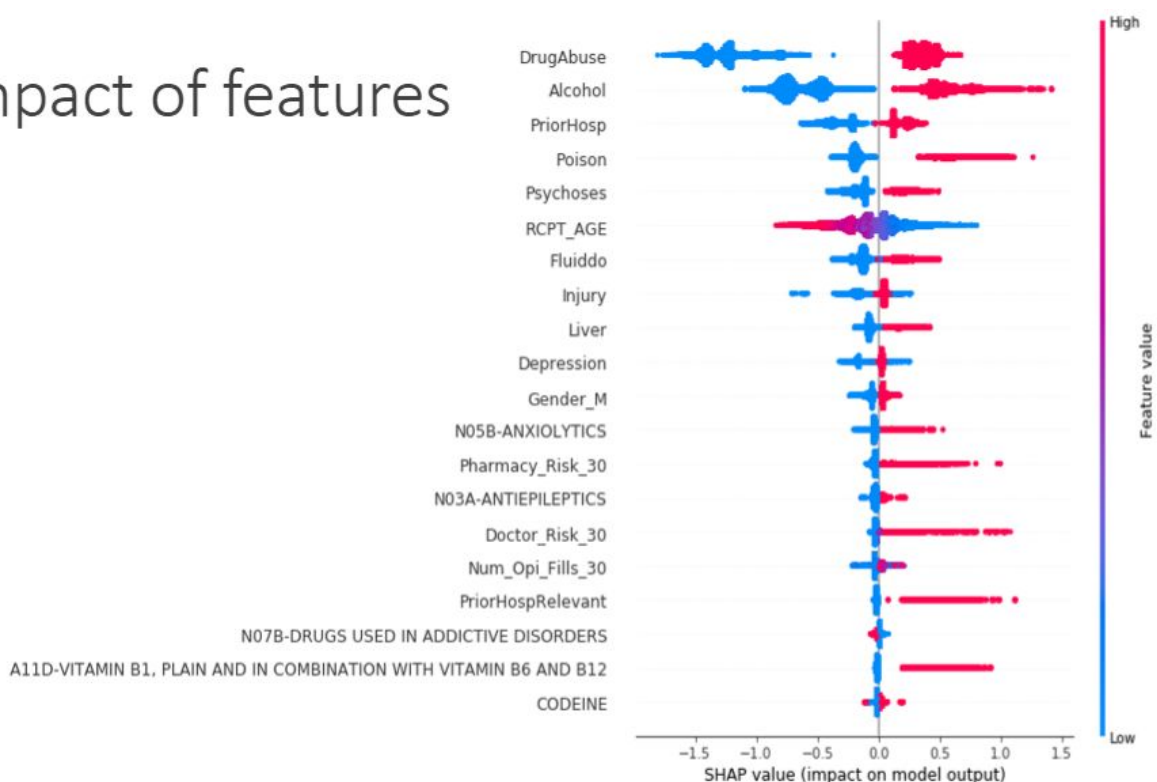
eFigure 3. Shapley values and feature impact in the XGBoost classifier using the 2018 validation set to describe “associations” between features and the outcome. Features with the most impact on the model with drug abuse with drug abuse ranked highest (A); tornado plot illustrating feature impact. Red 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)



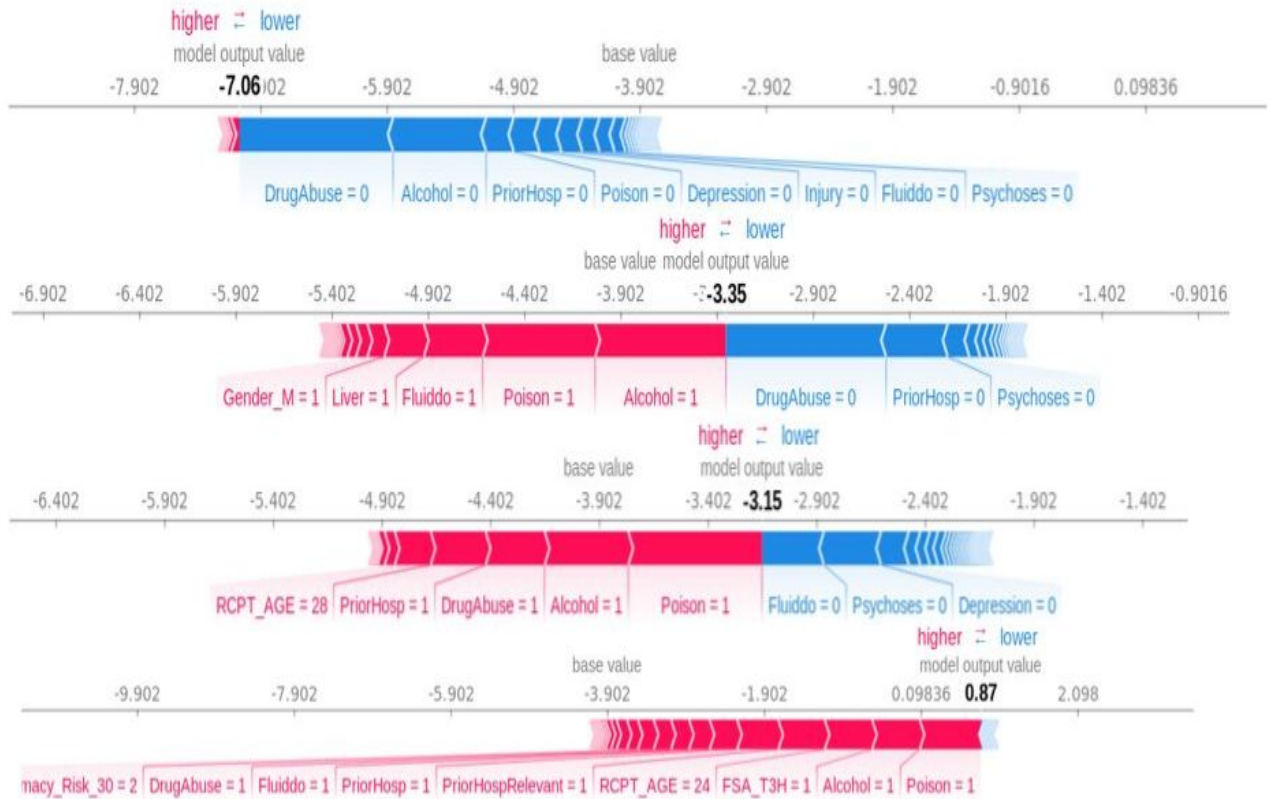
(B)

Impact of features



Note: RCPT_AGE- age at opioid dispensation; Fluiddo- fluid disorder according to Elixhauser co-morbidity; Gender_M-male sex' N05B-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:	<i>BMJ Open</i>
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 Kulkarni, Vinaykumar; OKAKI Health Analytics Eurich, Dean; University of Alberta, School of Public Health Kumar, Luke; Alberta Machine Intelligence Institute 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

SCHOLARONE™
Manuscripts



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.

1
2
3 **Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk**
4 **after an opioid dispensation in Alberta, Canada**
5

6 **Author list (in order):**
7

8 Vishal Sharma (0000-0001-7907-1183), Vinaykumar Kulkarni, Dean T. Eurich (0000-0003-2197-
9 0463), Luke Kumar, Salim Samanani (0000-0001-6751-4805)
10
11

12
13
14 **Address for each author:**
15

16 2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of
17 Alberta, Edmonton, Alberta, Canada, T6G 2E1 Vishal Sharma BPharm PhD Candidate,
18
19

20
21 OKAKI Health Intelligence, Edmonton, Alberta, Canada, Vinaykumar Kulkarni MSc
22
23

24
25 2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of
26 Alberta, Edmonton, Alberta, Canada, T6G 2E1 Dean Eurich professor
27
28

29
30 Alberta Machine Intelligence Institute, Edmonton, Alberta, Canada, T5J 3B1 Luke Kumar MSc
31
32

33
34 OKAKI Health Intelligence, Calgary, Alberta, Canada, Salim Samanani MD, Medical Director
35
36

37
38 **Corresponding Author:**
39

40 Dean Eurich, 2-040 Li Ka Shing Center for Health Research Innovation, University of Alberta,
41 Edmonton, Alberta, Canada, T6G 2E1; Phone 780-492-6333; fax 780-492-7455; email:
42 deurich@ualberta.ca
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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.

Copyright/license for publication: *The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide licence to the Publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) licence any third party to do any or all of the above.*

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).

1
2
3 **Data Sharing:** The data used in this study is not available for external analysis. However, administrative
4 health data can be accessed from Alberta Health by following defined research protocols and
5 confidentiality agreements.
6
7

8
9 **Transparency:** The lead author, VS, (the manuscript's guarantor, Dean Eurich) affirms that the
10 manuscript is an honest, accurate, and transparent account of the study being reported; that no
11 important aspects of the study have been omitted; and that any discrepancies from the study as
12 originally planned (and, if relevant, registered) have been explained.
13
14

15
16 **Word Count: 3120**
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

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 crisis¹; a major part of the policy response to the opioid crisis focuses on endorsing safe, appropriate opioid prescribing²⁻⁴. 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 ICD⁸ 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 populations¹¹. 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.

1
2
3 Supervised machine learning (ML)^{12,13} is an approach that uses computer algorithms to
4
5 build predictive models in the clinical setting that can make use of the large amounts of
6
7 available administrative data^{14,15}, all within a well-defined process¹⁶. Supervised ML trains on
8
9 labelled data to develop prediction models that are specific to different populations and, in
10
11 many cases, can provide better predictive performance than traditional, population-based
12
13 statistical models^{10,15,17}. We identified one study¹⁰ that applied ML techniques to predict
14
15 overdose risk in opioid patients pursuant to a prescription. In their validation sample, they
16
17 found that the DNN (deep neural network) and GBM (gradient boosting machine) algorithms
18
19 carried the best discrimination performance based on estimated c-statistics and that the ML
20
21 approach out-performed the guideline approach in terms of risk prediction; neural networks
22
23 have little interpretability and are not necessarily better at predicting outcomes when trained
24
25 on structured data¹⁸. This study relied on c-statistics to evaluate their ML models and did not
26
27 emphasize other performance metrics required to assess clinical utility that are recommended
28
29 by medical reporting guidelines^{11,13,19,20}. It also did not address the important issue of ML
30
31 model interpretability²¹. Reporting informative prognostic metrics is needed to better
32
33 understand the capabilities of ML classifiers if health departments and PMPs are to incorporate
34
35 them into their decision-making processes.
36
37
38
39
40
41
42
43
44

45 The objective of our study was to further develop and validate ML algorithms (beyond
46
47 just DNN) to predict the 30-day risk of hospitalization, emergency visit and mortality for a
48
49 patient in Alberta, Canada at the time of an opioid dispensation using administrative data
50
51 routinely available to health departments and PMPs and evaluate them using the above
52
53 referenced reporting guidelines. We also analyzed feature importance to provide meaningful
54
55
56
57
58
59
60

1
2
3 interpretations of the ML models. Comparing discrimination performance (area under the
4 receiver operating characteristics curves), we hypothesized that the ML process would perform
5 better than the current guideline approach for predicting risk of adverse outcomes related to
6 opioid prescribing.
7
8
9
10
11
12

13 **Methods**

14 **Study Design and Participants**

15
16
17 This prognostic study used a supervised ML scheme. All patients in Alberta, Canada who
18 received a dispensation for an opioid, were 18 years of age and older between Jan 1, 2017 and
19 Dec 31, 2018 were eligible. Patients were excluded from all analyses if they had any previous
20 diagnosis of cancer, received palliative interventions or were pregnant during the study period
21 (eTable 1 in Supplement) as use of opioids in these contexts is clinically different.
22
23
24
25
26
27
28
29

30
31
32 Government health departments and payers in many jurisdictions have systems to capture
33 prescription histories and ICD diagnostic codes. As such, we linked various administrative
34 health data sets available in Alberta, Canada using unique patient identifiers in order to
35 establish a complete description of patient demographics, drug exposures and health
36 outcomes. These databases include 1) *Pharmaceutical Information Network (PIN)*: PIN data
37 includes all dispensing records from community pharmacies from all prescriber types occurring
38 in the province outside of the hospital setting. PIN collects all drug dispensations irrespective of
39 age or insurance status in Alberta, 2) *Population and Vital Statistics Data (VS, Alberta Services)*:
40 sex, age, date of birth, death date, immigration and emigration data, and underlying cause of
41 death according to the World Health Organization algorithm using ICD codes⁸, 3)
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57

1
2
3 *Hospitalizations and Emergency Department Visits (NACRS [National Ambulatory Care*
4 *Reporting System], DAD [Discharge Abstract Database]): all services, length of stay, diagnosis*
5
6 (up to 25 ICD-10⁸ based diagnoses). Data and coding accuracy are routinely validated both
7
8 provincially and centrally via the Canadian Institute for Health Information, and 4) *Physician*
9
10 *Visits/Claims (Alberta Health): all claims from all settings (e.g., outpatient, office visits,*
11
12 *emergency departments, inpatient) with associated date of service, ICD code, procedure and*
13
14 *billing information.*
15
16
17
18
19
20

21 This study followed the TRIPOD and STARD reporting guidelines²²⁻²⁴ and received ethics
22
23 approval from the University of Alberta ethics board (Pro00083807_AME2). All analyses were
24
25 done using Python (v. 3.6.8.), SciKit Learn²⁵ (v. 0.23.2) SHAP²⁶ (v. 0.35), XGBoost (v. 0.90)²⁷,
26
27 Pandas (v. 1.0.5)²⁸ and H2O Driverless AI (version 1.9).
28
29
30

31 **Measures and Outcome**

32
33
34 ML models were trained on a labelled dataset in which the observation/analysis unit was an
35
36 opioid dispensation. The primary outcome was a composite of a drug-related hospitalization,
37
38 emergency department (ED) visit or mortality within 30 days of an opioid dispensation based on
39
40 ICD-10 codes (T40, F55, F10-19; eTable 2 in Supplement)^{2,10,29}.
41
42
43
44

45 We anticipated that our defined outcome would be a rare event, leading to a class
46
47 imbalanced dataset³⁰. To address this, we relied on specifying balanced class weightage for
48
49 supporting algorithms; other approaches were not deemed suitable (e.g., randomly repeating
50
51 minority class) and under sampling (sub-sampling within the majority class) resulted in changes
52
53 in outcome prevalence.
54
55
56
57

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 codes³¹), co-morbidity history using ICD-based Elixhauser score categories³², health care utilization (number of unique opioid prescribers, number of hospital visits), and drug utilization (level 3 ATC codes³³, oral morphine equivalents³⁴, 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%) sets¹³ 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

1
2
3 who did not, using chi-squared tests of independence. As well, we used all 2018 data as
4
5 another independent validation set.
6

7
8
9 First, we trained commonly used^{13,35} ML algorithms (eAppendix in Supplement) and out
10
11 of box models were further tuned when training on the dataset using 5-fold cross validation on
12
13 the training data. to address model overfitting^{13,36}. As is common in ML validation studies^{10,13},
14
15 we reported model discrimination performance (i.e. how well a model differentiates those at
16
17 higher risk from those at lower risk)¹¹ using area under the receiver operating characteristic
18
19 curve (AUROC; c-statistic). We then stratified the two ML models with the highest c-statistics
20
21 into percentile categories according to absolute risk of our outcome, as was done in previous
22
23 studies^{10,37}. We also plotted AUROC¹¹ and precision-recall curves (PRCs)³⁸.
24
25
26
27
28

29 Because discrimination alone is insufficient to assess ML model prediction capability, we
30
31 assessed a second necessary property, namely, calibration (i.e., how similar the predicted
32
33 absolute risk is to the observed risk across different risk strata)^{11,39}. Using the two ML models
34
35 with the highest discrimination performance discussed above, we assessed calibration
36
37 performance on the 2018 data by plotting observed (fraction of positives) vs predicted risk
38
39 (mean predicted value). Using these two ML classifiers, we analyzed the top 0.1, 1, 5, and 10
40
41 percentiles of predicted risk by the number of true and false positives, positive likelihood ratios
42
43 (PLR)²⁰, post-test probabilities, and number needed to screen. We also performed a simulation
44
45 of daily data uploads for 2018 Quarter 1 to view the predictive capabilities if a ML risk predictor
46
47 were to be deployed into a monitoring workflow.
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 For the XGBoost and logistic regression classifiers, we reported feature importance³⁵
4
5 and plotted PRCs that compared all dispenses to those within the top 10 percentiles of
6
7 estimated risk. As well, for the XGBoost classifier, we described feature importance on model
8
9 outcome using SHAP values^{26,40} to add an additional layer of interpretability.
10
11

12
13 Finally, we compared ML risk prediction (the two ML models with highest discrimination
14
15 performance) to current guideline approaches as others have¹⁰, using the 2019 Centers for
16
17 Medicare & Medicaid Services (CMS) opioid safety measures⁴¹ and the 2017 Canadian Opioid
18
19 Prescribing Guideline³. We also compared the discrimination performance of different logistic
20
21 regression classifier models using various combinations of features derived from their
22
23 respective databases: **1)** demographic and drug/health utilization features from PIN and **2)** co-
24
25 morbidity features derived from DAD, NACRS and Claims.
26
27
28
29

30 31 **Patient and Public Involvement**

32
33
34 This research was done without patient involvement. Patients were not invited to comment on
35
36 the study design and were not consulted to develop patient relevant outcomes or interpret the
37
38 results. Patients were not invited to contribute to the writing or editing of this document for
39
40 readability or accuracy. There are no plans to disseminate the results of the research to study
41
42 participants.
43
44
45

46 47 **Results**

48 49 **Patient Characteristics and Predictors**

50
51
52
53
54
55
56
57
58
59
60

1
2
3 We identified 392,979 patients with at least one opioid dispensation in 2017 (Figure 1). This
4 cohort was used to train (n= 275,150, 70%) and validate (n=117,829, 30%) ML models. In 2017
5 and 2018, 6,608 and 5,423 patients experienced the defined outcome, respectively. Baseline
6 characteristics were different between those who experienced the outcome and those who did
7 not (eTable 3 in Supplement) while characteristics were similar between the training and
8 validation sets (eTable 4 in Supplement). There were 2,283,075 opioid dispensations in 2017
9 and 1,977,389 in 2018. Overall, in 2017, 2.03% (n= 45,757) of opioid dispensations were
10 associated with the outcome; in 2018, the estimate was 1.6% (n= 31,392).
11
12
13
14
15
16
17
18
19
20
21
22

23 As described above, we categorized our candidate features into four groups (eTable 5 in
24 Supplement).
25
26
27

28 **Machine-Learning Prediction Performance**

29
30
31 Using the 2017 validation set, AUROCs for the XGBoost and logistic regression classifiers had
32 the highest discrimination performance at 0.87, while the neural network classifier had lower
33 performance at 0.80 (eTable 6 in Supplement).
34
35
36
37
38
39

40 Discrimination performance was similar for the 2018 validation set (n=393,023; eTable 6
41 in Supplement). XGBoost and logistic regression had the highest estimated AUROCs and area
42 under PRCs while the neural network classifier was lower (Figure 2A, 2B). As expected,
43 precision-recall curves indicate stronger predictive performance in opioid dispensations at
44 higher predicted risk percentiles (Figure 2C, 2D).
45
46
47
48
49
50
51

52
53 In the 2018 validation set, although discrimination performance was similar (0.88),
54 individual feature importance was different between the logistic regression and XGBoost
55
56
57

1
2
3 classifiers, with logistic regression feature importance more reliant on co-morbidity data from
4
5 DAD, NACRS and Claims while XGBoost relied more on drug utilization data from PIN (eFigure
6
7 2). In the XGBoost classifier, history of drug abuse, alcoholism, and prior hospitalization carried
8
9 the highest importance for predicting the study outcome (eFigure 3A) where the presence of
10
11 these features in a patient suggested a strong prediction towards having the defined outcome
12
13 (eFigure 3B and 3C).
14
15
16
17

18 **Calibration**

19
20
21 When considering dispensations predicted to be in the highest percentiles of risk, the top 5-
22
23 percentile captured 42% of all outcomes using the XGBoost and logistic regression classifiers
24
25 (Table 1). Also, as the predicted risk percentiles get higher (top 10 percentile to top 0.1
26
27 percentile), so too do the corresponding PPVs with the top 0.1 percentile associated with a PPV
28
29 of 33% for the XGBoost classifier. As well, lower categories of risk percentiles were associated
30
31 with lower outcomes (Figure 3, eFigure 4). When we simulated a monitoring workflow scenario
32
33 with daily data uploads, a similar pattern was illustrated where the dispensations predicted to
34
35 be higher risk had higher event rates (Figure 4).
36
37
38
39
40

41
42 After using the XGBoost and logistic regression classifiers to identify the dispensations in the
43
44 highest predicted risk percentiles, the pre-test probability of the outcome (1.6%) was
45
46 transformed into higher post-test probabilities, with higher probabilities in the riskier
47
48 percentiles (Table 1). The number needed to screen also decreased as predicted risk increased
49
50 (Table 1).
51
52
53
54
55
56
57
58
59
60

1
2
3 Comparing discrimination performance, ML risk prediction outperformed the current
4
5 guideline approaches when using various combinations of guideline recommendations (Table
6
7 2). In many of the guideline scenarios, the estimated AUROCs were close to the 0.5 mark.
8
9 When we estimated the discrimination performance of the logistic regression classifier based
10
11 on database source, using all databases produced an AUROC of 0.88. Reducing the database
12
13 source to only DAD, NACRS, Claims (co-morbidities only) resulted in an AUROC of 0.85, while
14
15 PIN (prescription history) only was 0.78 (Table 3).
16
17
18
19
20

21 Discussion

22
23
24 This study showed that ML techniques using available administrative data (prescription
25
26 histories and ICD codes) may provide enough discriminatory performance to predict adverse
27
28 outcomes associated with opioid prescribing. Indeed, our ML analyses showed very high
29
30 discrimination performance at 0.88. The linear model (logistic regression) and XGBoosted Trees
31
32 carried higher discrimination and calibration performance, while the neural network classifier
33
34 did not perform as well. By identifying the predicted top 5-10 percentile of absolute risk
35
36 pursuant to an opioid dispensation, we were able to capture approximately half of all outcomes
37
38 using ML methods. All ML models we trained had higher discrimination performance using
39
40 independent (external) validation sets than the clinical guideline approach.
41
42
43
44
45
46

47 Since the prevalence of our defined outcome is relatively low in the general population,
48
49 PPVs would also be expectedly low. However, estimated PPVs increased when we considered
50
51 higher risk dispensations, as is expected since PPV is related to event prevalence. This is
52
53 important because different users of a risk predictor will require different predictive
54
55
56
57
58
59
60

1
2
3 capabilities. Similarly, our estimates of positive likelihood ratios and associated post-test
4
5 probabilities also increased in dispensations with higher predicted risk indicating the strong
6
7 predictive capabilities of the XGBoost and logistic regression classifiers; likelihood ratios >10
8
9 generate conclusive changes from pre-test to post-test probabilities²⁰.
10
11

12
13 The current guideline approach to assess absolute opioid prescribing risk produced c-
14
15 statistic estimates closer to 0.5 indicating that discrimination was not much better than chance
16
17 alone. ML models with higher predictive performance can better support health departments
18
19 and PMPs with monitoring mandates to identify and intervene on those at high risk and their
20
21 associated prescribers. We also found that adding co-morbidity features from administrative
22
23 databases increased prediction performance compared to prescription history alone, thus
24
25 making the case for the use of this data by PMPs and health departments. However, if only
26
27 prescription history is available, our trained XGBoost classifier still had strong discrimination
28
29 performance.
30
31
32
33
34
35

36 We found only one study that used ML approaches to quantify the absolute risk of an
37
38 event pursuant to an opioid dispensation¹⁰. Their methodology used rolling 3-month windows
39
40 for estimating risk and ML model training while we used historic records to estimate 30-day
41
42 risk. Differences in study population and feature selection may explain why their highest
43
44 performing ML model was deep learning (neural network classifier) and ours was not.
45
46 Nevertheless, we were able to replicate their predictive performance using our ML approach as
47
48 we both showed that ML approaches have higher predictive capabilities than guideline
49
50 approaches. Both of our studies used predicted percentile risk estimates to identify high risk
51
52 dispensations and were able to do so with strong discrimination and calibration performance.
53
54
55
56
57

1
2
3 Furthermore, we emphasized prognostic metrics which are more informative to assess the
4 clinical utility of ML classifiers using pre- and post-test probabilities, something not done in
5 other studies and recommended in medical guidelines²⁰. This major aspect of our study, not
6 done previously, is important because any ML classifier that does not increase prognostic
7 information compared to baseline cannot be incorporated into decision making for the purpose
8 of intervening on higher risk instead of lower risk patients. Indeed, another study we found
9 describes how identifying cases in higher predicted risk percentiles using ML methods can be
10 deployed in hospital settings for the purpose of targeted interventions³⁷ upon discharge,
11 however the effect on outcomes is still to be determined.
12
13
14
15
16
17
18
19
20
21
22
23
24
25

26 The limitations of our study are similar to other ML studies¹⁰ and need to be addressed
27 when considering deployment of ML risk predictors. Our training dataset was not able to
28 account for non-prescription opioid consumption and the risk associated with non-prescription
29 use, both of which are substantial contributors to overall risk². Regarding our analysis, we
30 assumed that all dispensations were independent events; future research in this area should
31 focus on employing ML methods using correlated data. As with all ML projects, our models
32 were trained using Alberta data and might not be generalizable to other populations, or to
33 specific populations within Alberta. However, our analyses were done on a large population
34 and these results would be expected to be generalizable to the vast majority of patients.
35 Moreover, one of the benefits of the ML process is that models can be retrained or similar
36 methods could be used to develop new models to accommodate different populations.
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52

53 This study suggests that ML risk prediction can support PMPs, especially if readily
54 available administrative health data is used. PMPs currently use population-based guidelines
55
56
57

1
2
3 which we, and others, have shown cannot predict absolute individual risk. The ML process
4
5 allows for model training, validation and deployment to specific settings in which, for the case
6
7 of PMPs, high risk patients can be identified and targeted for intervention either at the patient
8
9 or provider level. Moreover, ML classifiers can be retrained over time as changes in
10
11 populations and trends in prescribing occur and are therefore specific to the population unlike
12
13 broadly based guidelines. Further research can assess whether implementation of a ML-based
14
15 monitoring system by PMPs leads to improved clinical outcomes.
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 Pain Opioid 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. Hsieh 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 mining 2015.
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 systems 2017.
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 mining 2016.
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.whooc.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.

1
2
3 *The Canadian guidelines do not specify timelines. >90 OME was determined by taking the average
4 daily OME over the 30 days prior to dispensation
5

6 **The CMS guidelines specify a timeline of 90 or more days at >120 OME and concurrent use of
7 opioids and benzodiazepines for 30 days or more
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

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 equivalents³⁴, 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

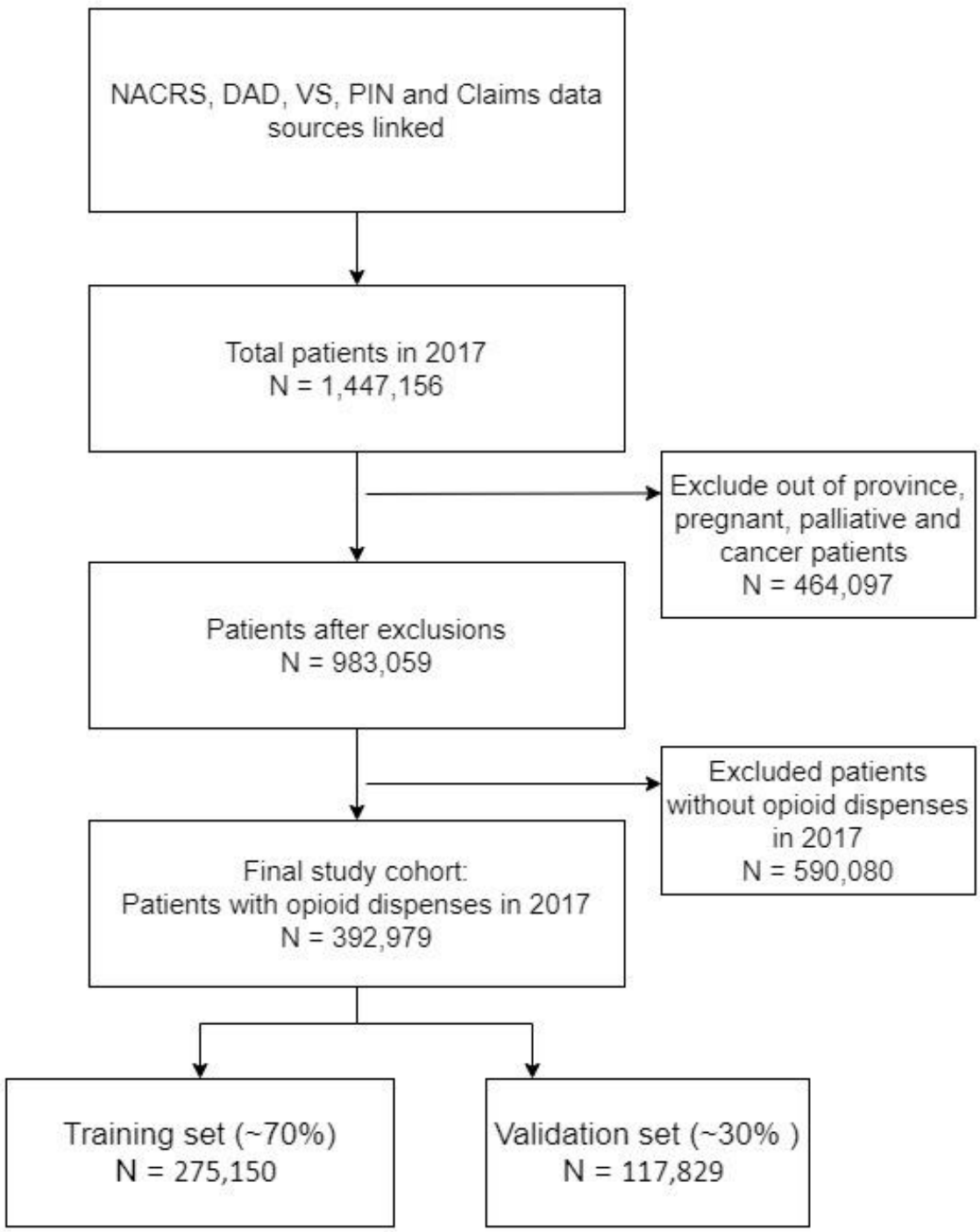
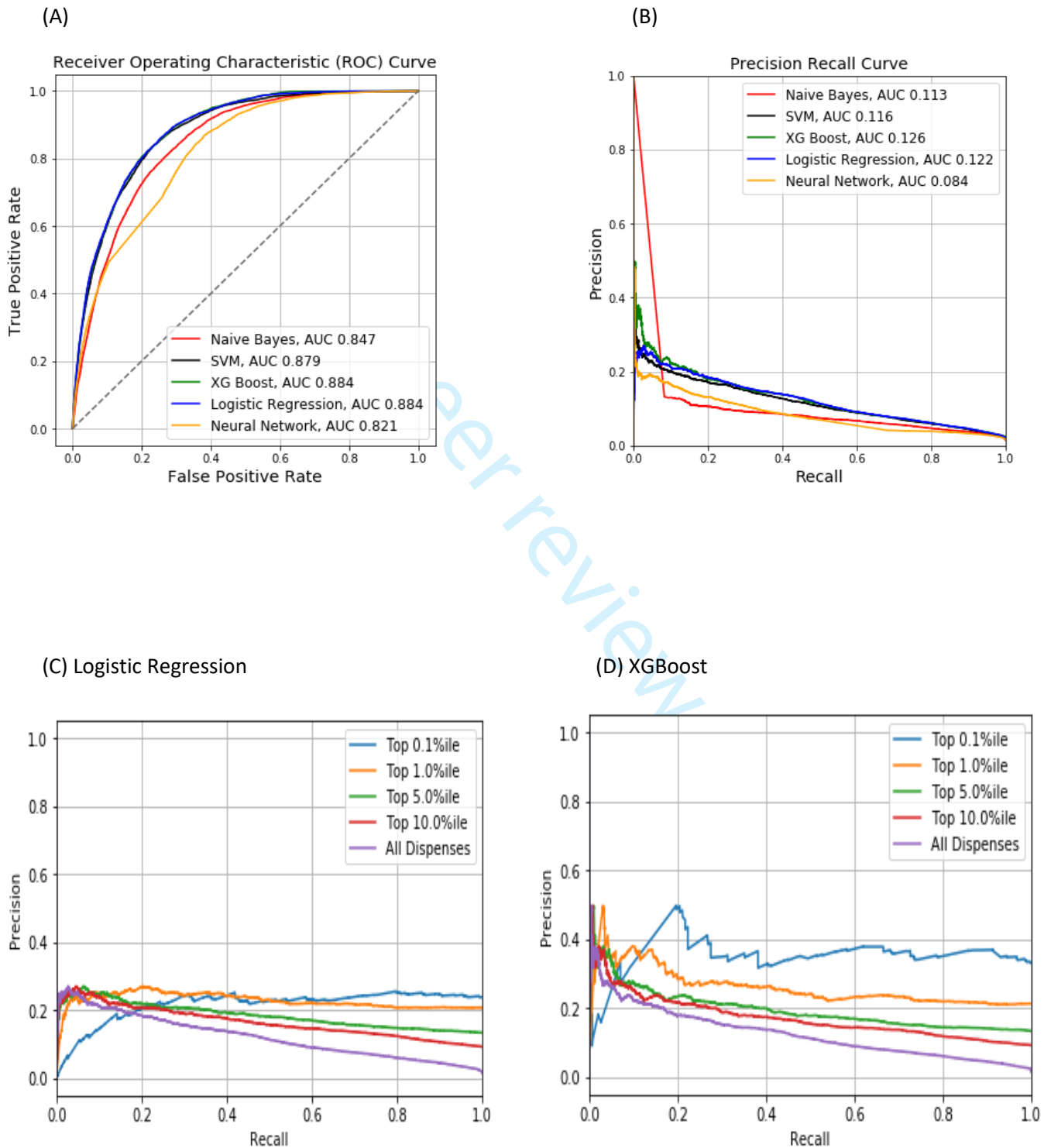


Figure 2. Area under the receiver operating characteristic curve (A) and precision-recall curves (B) for all dispensations using logistic regression (L1), neural network, support vector machine (SVM), XGBoost and Naïve-Bayes; precision-recall curves for higher risk dispensations according to predicted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



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

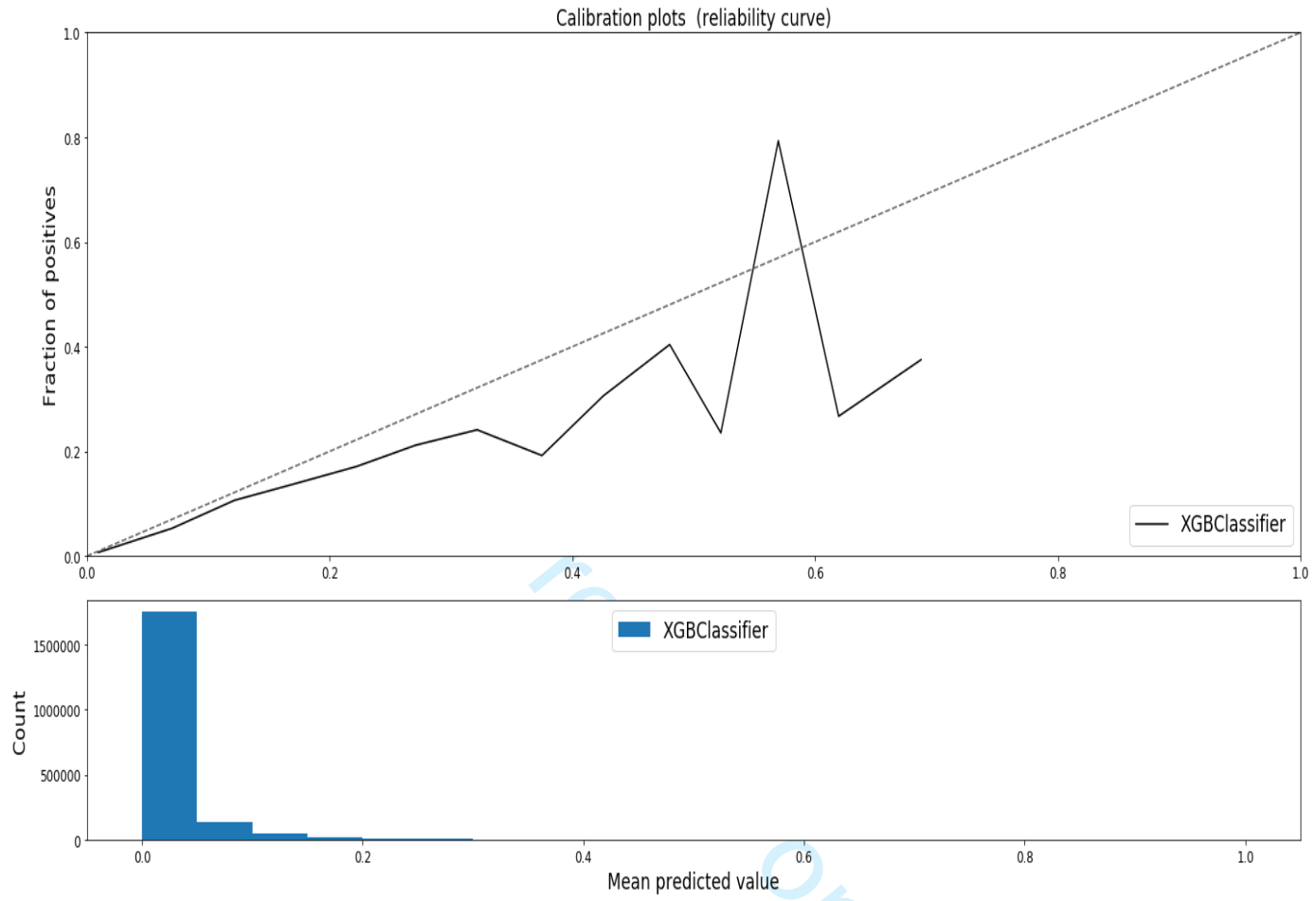
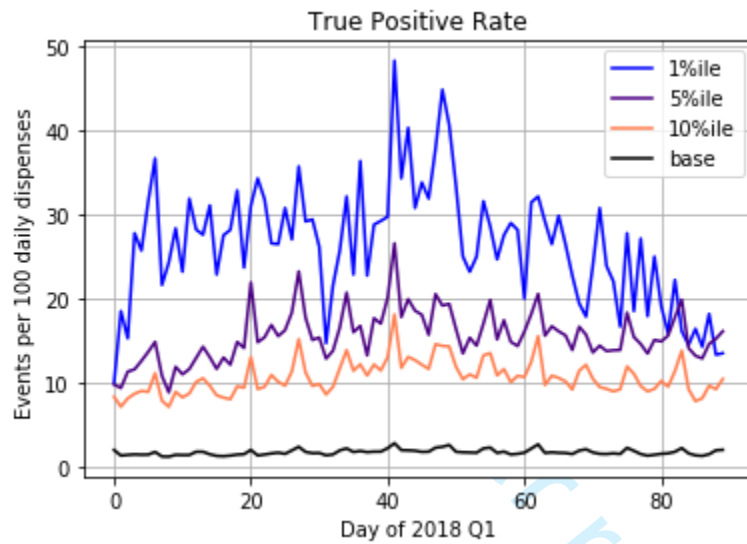
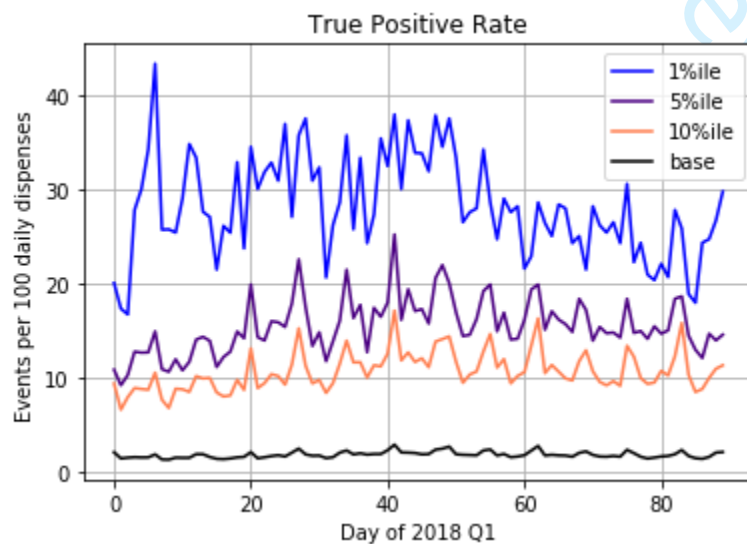


Figure 4. Simulation of a clinical workflow with daily uploads and events per 100 daily dispenses by risk percentiles using 2018 Quarter 1 (Q1) data for logistic regression (A) and XGBoost (B) classifiers. For both classifiers, the majority of counts (dispensations) were predicted to be lower risk.

(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 chi² 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

1
2
3 functions has shown to be very effective recently in many application domains such as natural
4 language processing, computer vision, genomics, computer games and health[2]. Neural
5 networks come in many flavors learning nonlinear mapping of different types of data such as
6 Convolutional NNs being most effective with images and Recurrent NNs for time series and
7 language data. Identifying the most effective neural network structure is one of the difficult and
8 the most time-consuming aspect of applying neural networks to new application domains and
9 data. Generally, neural networks try to exploit the relationships in the raw unstructured data (eg:
10 image and text) presented to the network but with more structured data such as health records
11 and ICD codes learning relationships is much complex. Our neural network models are mainly
12 based on densely connected hidden layers with ReLu[6] activation function. We used the cross-
13 entropy loss for the binary classification Adam optimizer. We used a simple feed forward
14 network using Sklearn MLP classifier with hyperparameter tuning for the NN.

19 **Boosted Learning Algorithms**

20
21 Boosting is a process to ensemble multiple base learning algorithms to arrive at better overall
22 performance than any individual base learner[1]. In contrast to independently building multiple
23 models from the subsets of the data, boosting re-weights the training data every time a model is
24 learned for future models. This weighting happens to give more preference to currently
25 misclassified data points in the next round compared to the correctly classified data points.
26 Therefore future learners try to do better on the misclassified data points leading to a collection
27 base learners having a better-combined prediction. This process is sequential so each base
28 learner is dependent on the output of the previously trained model (it is worthy to note XGBoost
29 provides a parallel tree boosting alternative). In our work, we have experimented with several
30 boosting meta-learning algorithms such as XGBoost[7], AdaBoost[5] and GBM[5]. XGBoost uses
31 a variant of trees as the base learner whereas AdaBoost (from Sci-kit learn) can use many ML
32 algorithms as base learners. GBM uses logistic regression by default as the base learner. We used
33 all 3 types of boosting with tuned hyperparameters for comparison.

39 **Naive Bayes**

40
41 Naive Bayes is based on the Bayes theorem with a strong independence assumption between the
42 covariates[1]. This assumption helps in building a simple probabilistic model for learning and
43 inference. Naive Bayes coefficients scale linearly with the number of covariates making this a
44 suitable model for high-dimensional data. We used Naive Bayes as a simple baseline learning
45 algorithm for comparison.

48 **Support Vector Machines (SVM)**

49
50 SVMs[4] are maximum margin classifiers optimizing for learning a hyperplane having the
51 maximum distance away from each of the class data points[1]. SVM is a linear classifier but with
52 the kernel trick to map the inputs to the higher dimensional space, it can learn nonlinear decision
53 boundaries in the input space. SVMs are very effective binary classifiers with the kernel trick[1].
54 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 χ^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

1
2
3 ^ Injury: ICD10: S00-T98; Poisoning: ICD10: T36-T50

4 ** Complicated, uncomplicated diabetes and hypertension were collapsed into one category each
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

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 chi² 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.

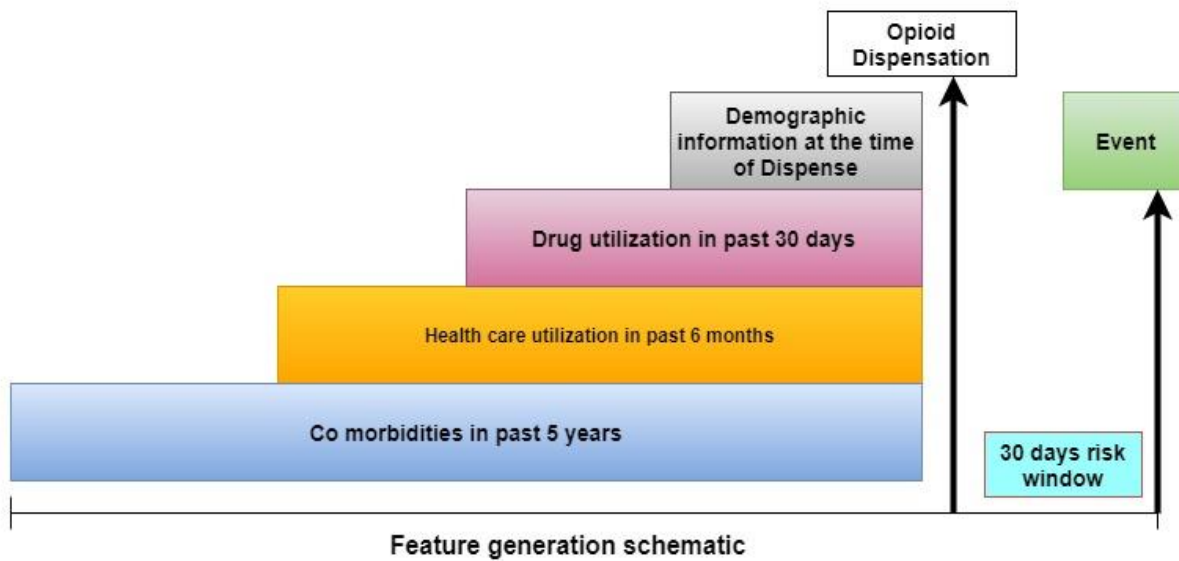
Note: ICD: International Statistical Classification of Diseases and Related Health Problems, World Health Organization.

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



review only

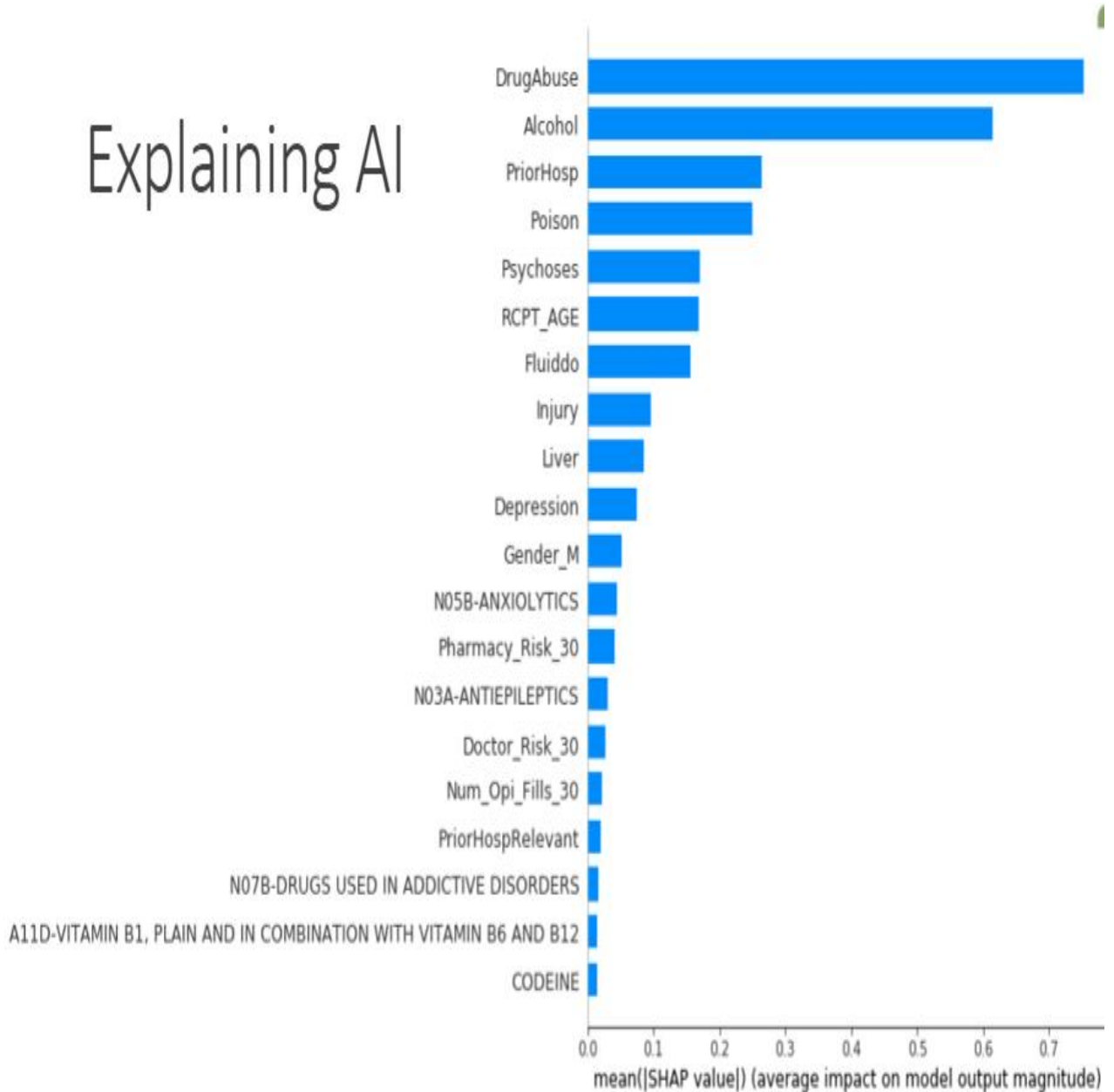
eFigure2. Feature importance from logistic regression and tree-based (XGBoost) classifiers using the 2018 validation set. The logistic regression classifier relied more on co-morbidity data from DAD, NACRS, and Claims databases; XGBoost classifier relied more on data from the PIN database. AUROCs for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

LOGISTIC REGRESSION		TREE-BASED MODEL	
Drug Abuse	1.00	Age	1.00
Age	0.65	Num_Fills_30	1.00
Prior Hospitalization	0.62	Num_Opioid_Fills_30	0.86
Alcohol Abuse	0.62	Num_Benzo_Fills_30	0.46
Fluid Disorder	0.32	Doctor_Risk_30	0.45
Substance Poison	0.31	Total_OME_30_Days_Supply	0.43
Psychoses	0.31	Substance poison	0.37
Num_Benzo_Ingred_30	0.26	Pharmacy_Risk_30	0.35
Depression	0.19	Num_Doctors_30	0.34
Concurrent_Opioid_Benzo_30	0.19	Income	0.34
Injury	0.17	Prior hospitalization	0.26

Note: Logistic regression used L1 (lasso) parameter regularization

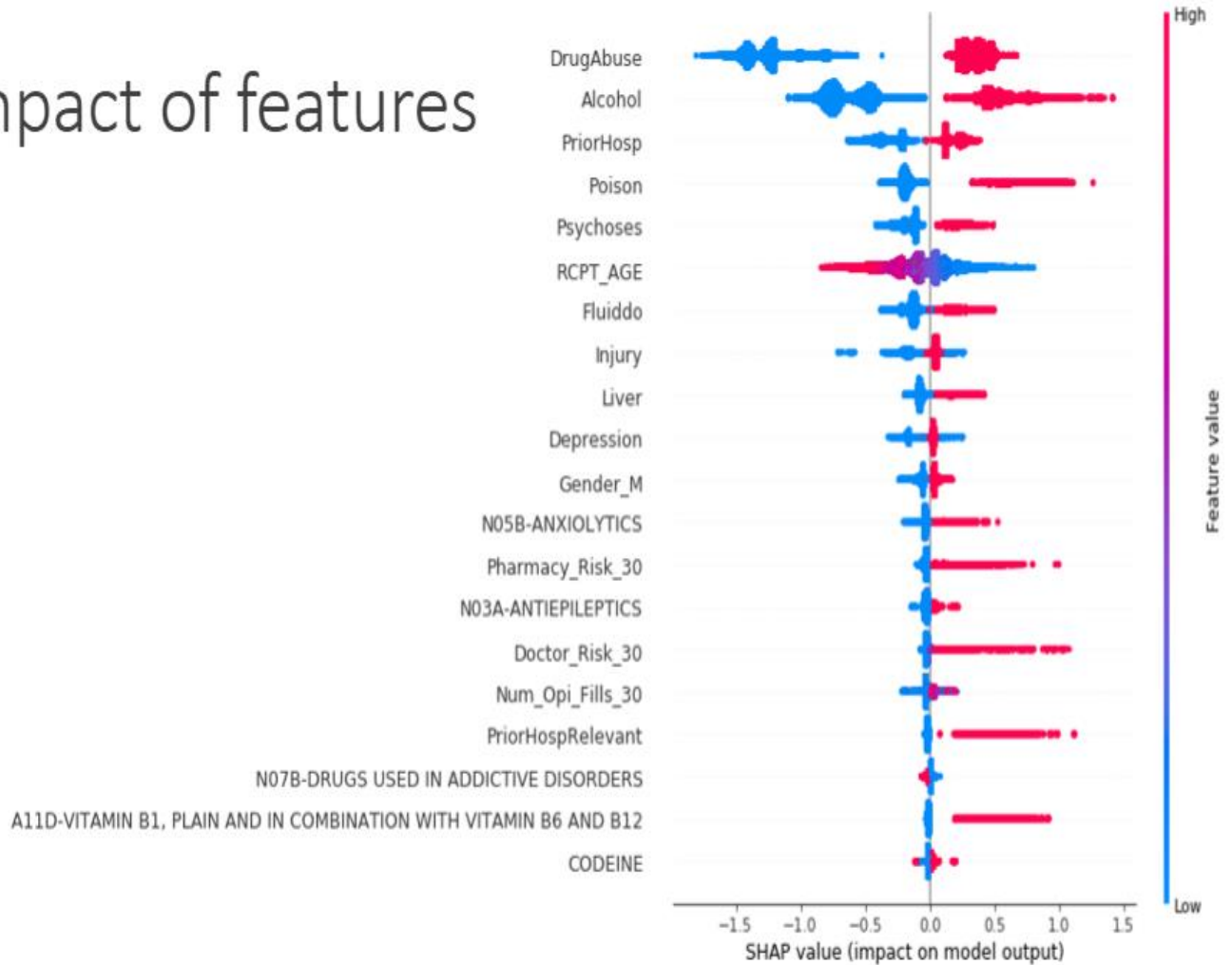
eFigure 3. Shapley values and feature impact in the XGBoost classifier using the 2018 validation set to describe “associations” between features and the outcome. Features with the most impact on the model with drug abuse with drug abuse ranked highest (A); tornado plot illustrating feature impact. Red 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)



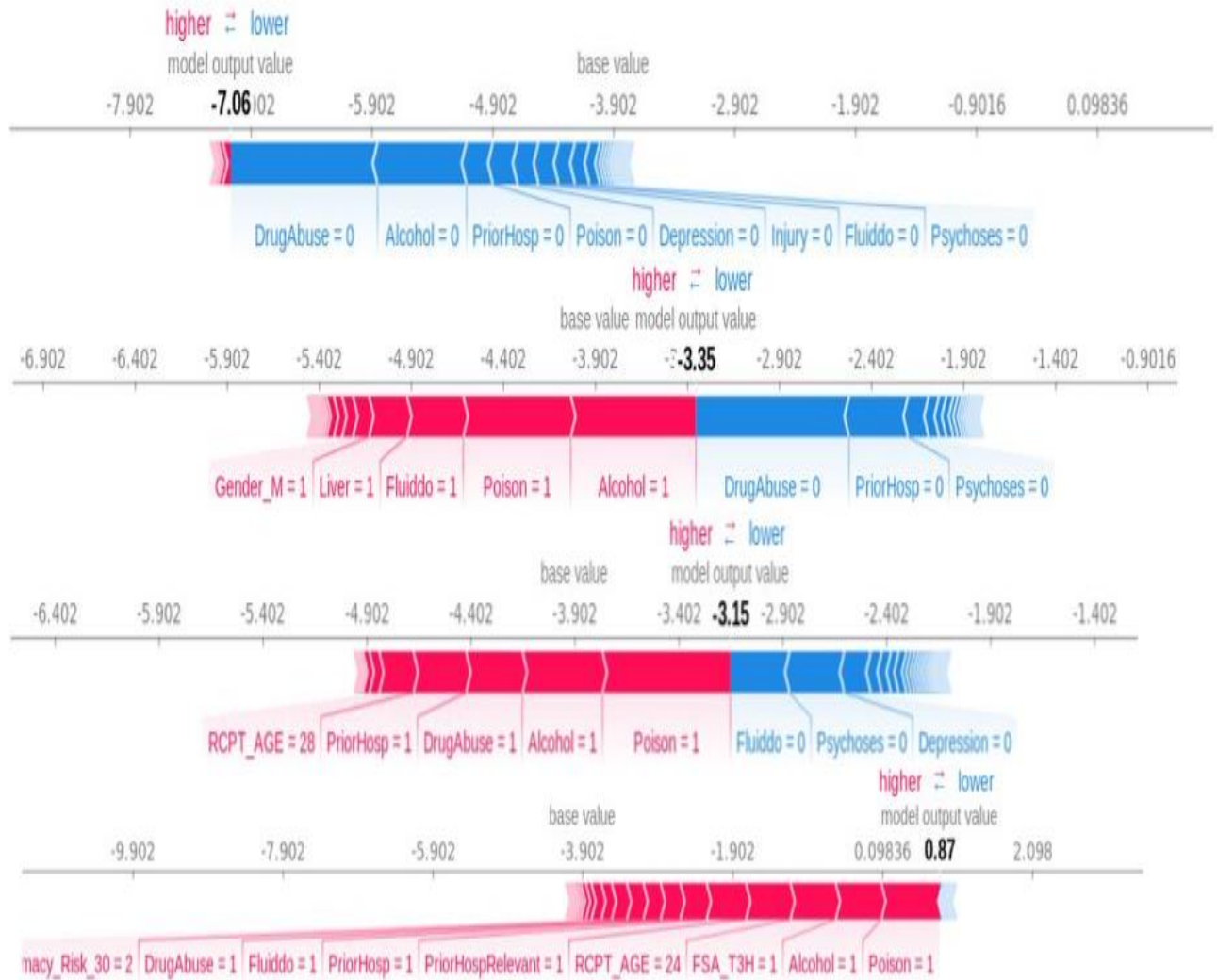
(B)

Impact of features



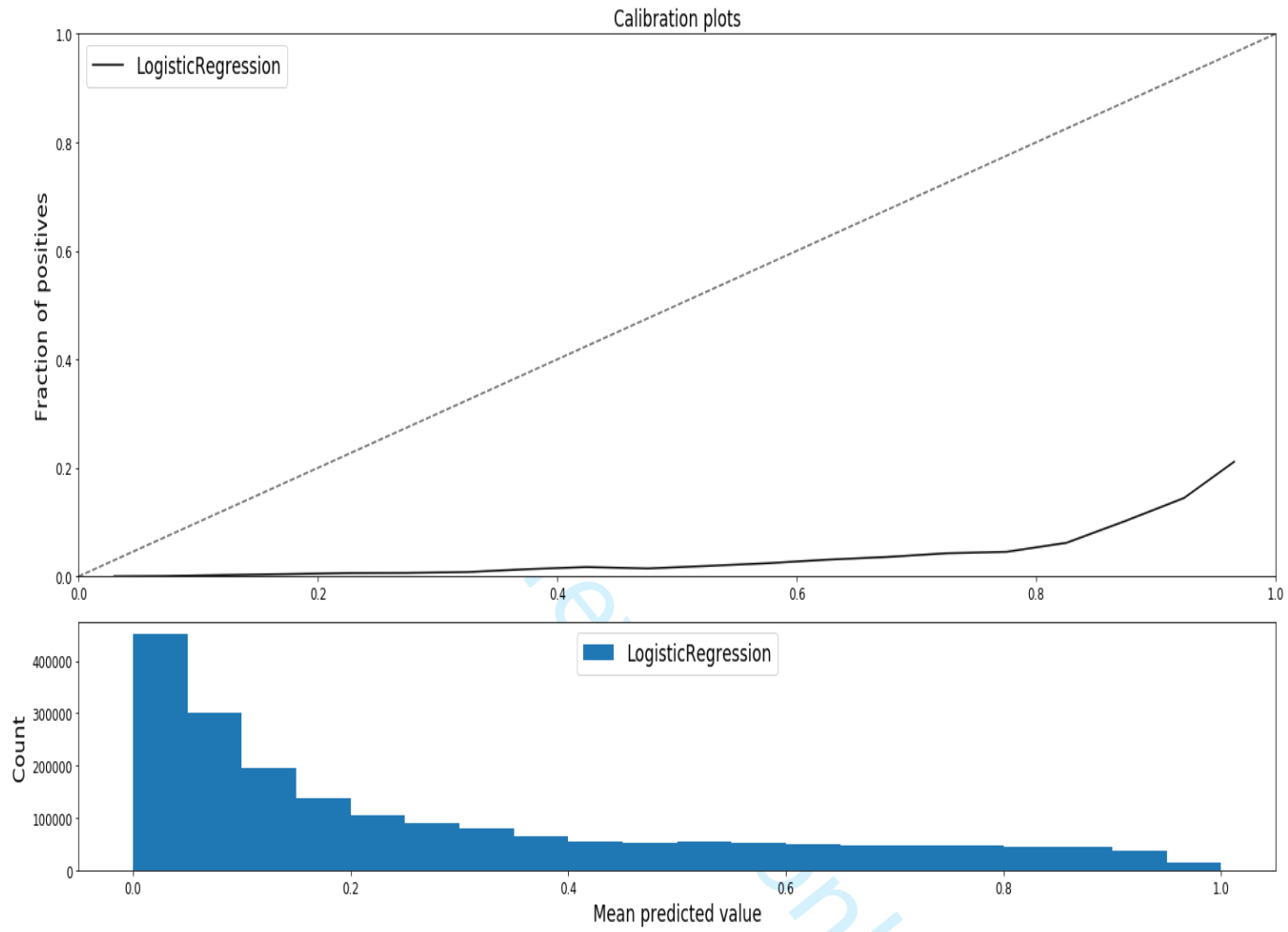
Note: RCPT_AGE- age at opioid dispensation; Fluiddo- fluid disorder according to Elixhauser co-morbidity; Gender_M-male sex' N05B-ANXIOLYTICS- prescribed ATC code benzodiazepine derivatives; Pharmacy_Risk_30- derived feature using proportion of opioid/benzodiazepine patients that experienced the study 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:	<i>BMJ Open</i>
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 Kulkarni, Vinaykumar; OKAKI Health Analytics Eurich, Dean; University of Alberta, School of Public Health Kumar, Luke; Alberta Machine Intelligence Institute 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

SCHOLARONE™
Manuscripts



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.

1
2
3 **Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk**
4 **after an opioid dispensation in Alberta, Canada**
5

6 **Author list (in order):**
7

8 Vishal Sharma (0000-0001-7907-1183), Vinaykumar Kulkarni, Dean T. Eurich (0000-0003-2197-
9 0463), Luke Kumar, Salim Samanani (0000-0001-6751-4805)
10
11

12
13
14 **Address for each author:**
15

16 2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of
17 Alberta, Edmonton, Alberta, Canada, T6G 2E1 Vishal Sharma BPharm PhD Candidate,
18
19

20
21 OKAKI Health Intelligence, Edmonton, Alberta, Canada, Vinaykumar Kulkarni MSc
22
23

24
25 2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of
26 Alberta, Edmonton, Alberta, Canada, T6G 2E1 Dean Eurich professor
27
28

29
30 Alberta Machine Intelligence Institute, Edmonton, Alberta, Canada, T5J 3B1 Luke Kumar MSc
31
32

33
34 OKAKI Health Intelligence, Calgary, Alberta, Canada, Salim Samanani MD, Medical Director
35
36

37
38 **Corresponding Author:**
39

40 Dean Eurich, 2-040 Li Ka Shing Center for Health Research Innovation, University of Alberta,
41 Edmonton, Alberta, Canada, T6G 2E1; Phone 780-492-6333; fax 780-492-7455; email:
42 deurich@ualberta.ca
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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.

Copyright/license for publication: *The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide licence to the Publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) licence any third party to do any or all of the above.*

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).

1
2
3 **Data Sharing:** The data used in this study is not available for external analysis. However, administrative
4 health data can be accessed from Alberta Health by following defined research protocols and
5 confidentiality agreements.
6
7

8
9 **Transparency:** The lead author, VS, (the manuscript's guarantor, Dean Eurich) affirms that the
10 manuscript is an honest, accurate, and transparent account of the study being reported; that no
11 important aspects of the study have been omitted; and that any discrepancies from the study as
12 originally planned (and, if relevant, registered) have been explained.
13
14

15
16 **Word Count: 3357**
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

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 crisis¹; a major part of the policy response to the opioid crisis focuses on endorsing safe, appropriate opioid prescribing²⁻⁴. 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 populations¹¹. 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.

1
2
3 Supervised machine learning (ML)^{12,13} is an approach that uses computer algorithms to
4
5 build predictive models in the clinical setting that can make use of the large amounts of
6
7 available administrative data^{14,15}, all within a well-defined process¹⁶. Supervised ML trains on
8
9 labelled data to develop prediction models that are specific to different populations and, in
10
11 many cases, can provide better predictive performance than traditional, population-based
12
13 statistical models^{10,15,17}. We identified one study¹⁰ that applied ML techniques to predict
14
15 overdose risk in opioid patients pursuant to a prescription. In their validation sample, they
16
17 found that the deep neural network (DNN) and gradient boosting machines (GBM) algorithms
18
19 carried the best discrimination performance based on estimated c-statistics and that the ML
20
21 approach out-performed the guideline approach in terms of risk prediction; neural networks
22
23 have little interpretability and are not necessarily better at predicting outcomes when trained
24
25 on structured data¹⁸. This study relied on c-statistics to evaluate their ML models and did not
26
27 emphasize other performance metrics required to assess clinical utility that are recommended
28
29 by medical reporting guidelines^{11,13,19,20}. It also did not address the important issue of ML
30
31 model interpretability²¹. Reporting informative prognostic metrics is needed to better
32
33 understand the capabilities of ML classifiers if health departments and PMPs are to incorporate
34
35 them into their decision-making processes.
36
37
38
39
40
41
42
43
44

45 The objective of our study was to further develop and validate ML algorithms (beyond
46
47 just DNN) to predict the 30-day risk of hospitalization, emergency visit and mortality for a
48
49 patient in Alberta, Canada at the time of an opioid dispensation using administrative data
50
51 routinely available to health departments and PMPs and evaluate them using the above
52
53 referenced reporting guidelines. We also analyzed feature importance to provide meaningful
54
55
56
57
58
59
60

1
2
3 interpretations of the ML models. Comparing discrimination performance (area under the
4 receiver operating characteristics curves), we hypothesized that the ML process would perform
5 better than the current guideline approach for predicting risk of adverse outcomes related to
6 opioid prescribing.
7
8
9
10
11
12

13 **Methods**

14 **Study Design and Participants**

15
16
17 This prognostic study used a supervised ML scheme. All patients in Alberta, Canada who
18 received a dispensation for an opioid, were 18 years of age and older between Jan 1, 2017 and
19 Dec 31, 2018 were eligible. Patients were excluded from all analyses if they had any previous
20 diagnosis of cancer, received palliative interventions or were pregnant during the study period
21 (eTable 1 in Supplement) as use of opioids in these contexts is clinically different.
22
23
24
25
26
27
28
29
30
31

32
33 Government health departments and payers in many jurisdictions have systems to
34 capture prescription histories and ICD diagnostic codes. As such, we linked various
35 administrative health data sets available in Alberta, Canada using unique patient identifiers in
36 order to establish a complete description of patient demographics, drug exposures and health
37 outcomes. These databases include 1) *Pharmaceutical Information Network (PIN)*: PIN data
38 includes all dispensing records from community pharmacies from all prescriber types occurring
39 in the province outside of the hospital setting. PIN collects all drug dispensations irrespective of
40 age or insurance status in Alberta; Anatomical Therapeutic Chemical classification (ATC) codes²²
41 were used to identify opioid dispensations (eSupplement), 2) *Population and Vital Statistics*
42 *Data (VS, Alberta Services)*: sex, age, date of birth, death date, immigration and emigration
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 data, and underlying cause of death according to the World Health Organization algorithm
4
5 using ICD codes⁸, 3) *Hospitalizations and Emergency Department Visits (National Ambulatory*
6
7 *Care Reporting System [NACRS], Discharge Abstract Database [DAD])*: all services, length of
8
9 stay, diagnosis (up to 25 ICD-10⁸ based diagnoses). Data and coding accuracy are routinely
10
11 validated both provincially and centrally via the Canadian Institute for Health Information, and
12
13
14
15 4) *Physician Visits/Claims (Alberta Health)*: all claims from all settings (e.g., outpatient, office
16
17 visits, emergency departments, inpatient) with associated date of service, ICD code, procedure
18
19 and billing information.
20
21
22

23 This study followed the TRIPOD and STARD reporting guidelines²³⁻²⁵ and received ethics
24
25 approval from the University of Alberta ethics board (Pro00083807_AME1). All analyses were
26
27 done using Python (v. 3.6.8), SciKit Learn²⁶ (v. 0.23.2) SHAP²⁷ (v. 0.35), XGBoost (v. 0.90)²⁸,
28
29 Pandas (v. 1.0.5)²⁹ and H2O Driverless AI (version 1.9).
30
31
32

33 34 **Measures and Outcome**

35
36
37 ML models were trained on a labelled dataset in which the observation/analysis unit
38
39 was an opioid dispensation. Every opioid dispensation, not just the incident one, was used as a
40
41 potential instance to predict the risk of our outcome. The primary outcome was a composite of
42
43 a drug-related hospitalization, emergency department (ED) visit or mortality within 30 days of
44
45 an opioid dispensation based on ICD-10 codes identified from DAD, NACRS and Vital Statistics
46
47 (T40, F55, F10-19; eTable 2 in Supplement)^{2,10,30}.
48
49
50

51
52 We anticipated that our defined outcome would be a rare event, leading to a class
53
54 imbalanced dataset³¹. To address this, we relied on specifying balanced class weightage for
55
56
57

1
2
3 supporting algorithms; other approaches were deemed not suitable (e.g., oversampling using
4 randomly repeating minority class); under sampling (sub-sampling within the majority class)
5
6 resulted in changes in outcome prevalence. Class weightage is a commonly used method³² to
7
8 address class imbalance along with over and under-sampling approaches. However,
9
10 oversampling, which involves generating new opioid dispensations from the original data
11
12 distribution and is prone to introducing bias, is difficult due to the categorical nature of the data
13
14 and beyond the scope of this study. With under-sampling, which takes samples from the
15
16 majority class (in this case, no 30-day event after dispensation), we would not be able to use all
17
18 of the information provided by the data in instances with no outcome. Hence, we decided to
19
20 use the class weightage method which does not alter the data distribution. Instead, the
21
22 learning process is adjusted in a way that increases the importance of the positive class
23
24 (instances that led to a 30-day event)³³.
25
26
27
28
29
30
31
32

33 **Predictor Candidates for ML Models**

34
35
36 Predictor variables in our ML models included those that were informed by the
37
38 literature^{3,4,10} and those directly obtained from the data sets. These included features based on
39
40 demographics (age, sex, income using Forward Sortation index from postal codes³⁴), co-
41
42 morbidity history using ICD-based Elixhauser score categories³⁵, health care utilization (number
43
44 of unique providers, number of hospital and emergency department visits), and drug utilization
45
46 (level 3 ATC codes²², oral morphine equivalents³⁶, concurrent use with benzodiazepines,
47
48 number of opioid and benzodiazepine dispensations, number of unique opioid and
49
50 benzodiazepine molecules). Depending on the potential predictor and data availability, we
51
52 used data from 30 days to 5 years before the opioid dispensation to generate model features
53
54
55
56
57

1
2
3 (eFigure 1 in Supplement); 30 days was used to reflect the immediate nature of the risk and 5
4
5 years to fully capture co-morbidities. This approach aligns with how health providers would
6
7 assess patients using the entire history of co-morbidities and then the more immediate factors
8
9 in deciding on the need for a therapeutic as well as risk in patients. We performed experiments
10
11 to identify the features and data sets that contributed most to predicting the outcomes with a
12
13 view to minimizing the potential future data requirements for health departments and PMPs.
14
15
16
17

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

19
20
21 We randomly divided the patients in the 2017 portion of our study cohort into training
22
23 (70%) and validation (30%) sets¹³ by patients and opioid dispensations such that no patients in
24
25 the training set were in the validation set. Baseline characteristics and event rates were
26
27 compared in the training vs validation group, and between those who experienced the outcome
28
29 and those who did not using chi-squared tests of independence. As well, we used all the 2018
30
31 data as another independent validation set.
32
33
34
35

36
37 We trained commonly used^{13,37} ML algorithms (eAppendix in Supplement) and further
38
39 tuned out-of-box models using 5-fold cross validation on the training data to address model
40
41 overfitting^{13,38}. As is common in ML validation studies^{10,13}, we reported model discrimination
42
43 performance (i.e. how well a model differentiates those at higher risk from those at lower
44
45 risk)¹¹ using area under the receiver operating characteristic curve (AUROC; c-statistic). We
46
47 then stratified the two ML models with the highest c-statistics into percentile categories
48
49 (deciles) according to absolute risk of our outcome, as was done in previous studies^{10,39}. We
50
51 also plotted AUROC¹¹ and precision-recall curves (PRCs)⁴⁰.
52
53
54
55
56
57

1
2
3 Because discrimination alone is insufficient to assess ML model prediction capability, we
4
5 assessed a second necessary property, namely, calibration (i.e., how similar the predicted
6
7 absolute risk is to the observed risk across different risk strata)^{11,41}. Using the two ML models
8
9 with the highest discrimination performance, we assessed calibration performance on the 2018
10
11 data by plotting observed (fraction of positives) vs predicted risk (mean predicted value). Using
12
13 these same two ML classifiers, we analyzed the top 0.1, 1, 5, and 10 percentiles of predicted
14
15 risk by the number of true and false positives, positive likelihood ratios (PLR)²⁰, positive
16
17 predictive values (PPV), post-test probabilities, and number needed to screen. We also
18
19 performed a simulation of daily data uploads for 2018 Quarter 1 to view the predictive
20
21 capabilities if a ML risk predictor were to be deployed into a monitoring workflow.
22
23
24
25
26
27

28 For the XGBoost and logistic regression classifiers, we reported feature importance³⁷
29
30 and plotted PRCs that compared all dispenses to those within the top 10 percentiles of
31
32 estimated risk. As well, for the XGBoost classifier, we described feature importance on model
33
34 outcome using SHAP values^{27,42} to add an additional layer of interpretability.
35
36
37
38

39 Finally, we compared ML risk prediction (the two ML models with highest discrimination
40
41 performance) to current guideline approaches as others have¹⁰, using the 2019 Centers for
42
43 Medicare & Medicaid Services (CMS) opioid safety measures⁴³ and the 2017 Canadian Opioid
44
45 Prescribing Guideline³. We also compared the discrimination performance of different logistic
46
47 regression classifier models using various combinations of features derived from their
48
49 respective databases: **1)** demographic and drug/health utilization features from PIN and **2)** co-
50
51 morbidity features derived from DAD, NACRS and Claims.
52
53
54
55
56
57
58
59
60

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

1
2
3 Using the 2017 validation set, AUROCs for the XGBoost and logistic regression classifiers
4
5 had the highest discrimination performance at 0.87, while the neural network classifier had
6
7 lower performance at 0.80 (eTable 6 in Supplement).
8
9

10
11 Discrimination performance was similar for the 2018 validation set (n=393,023; eTable 6
12
13 in Supplement). XGBoost and logistic regression had the highest estimated AUROCs and area
14
15 under PRCs while the neural network classifier was lower (Figure 2A, 2B). As expected,
16
17 precision-recall curves indicate stronger predictive performance in opioid dispensations at
18
19 higher predicted risk percentiles (Figure 2C, 2D).
20
21
22

23
24 In the 2018 validation set, although discrimination performance was similar (0.88),
25
26 individual feature importance was different between the logistic regression and XGBoost
27
28 classifiers, with logistic regression feature importance more reliant on co-morbidity data from
29
30 DAD, NACRS and Claims while XGBoost relied more on drug utilization data from PIN (eFigure
31
32 2). With the XGBoost classifier, history of drug abuse, alcoholism, and prior
33
34 hospitalization/emergency visit carried the highest importance for predicting the study
35
36 outcome (eFigure 3A) where the presence of these features in a patient suggested a strong
37
38 prediction towards having the defined outcome (eFigure 3B and 3C).
39
40
41
42
43

44 **Calibration**

45
46
47 When considering dispensations predicted to be in the highest percentiles of risk, the
48
49 top 5-percentile captured 42% of all outcomes using the XGBoost and logistic regression
50
51 classifiers (Table 1). Also, as the predicted risk percentiles get higher (top 10 percentile to top
52
53 0.1 percentile), so too do the corresponding PPVs with the top 0.1 percentile associated with a
54
55
56
57
58
59
60

1
2
3 PPV of 33% for the XGBoost classifier. As well, lower categories of risk percentiles were
4
5 associated with lower outcomes (Figure 3, eFigure 4). When we simulated a monitoring
6
7 workflow scenario with daily data uploads, a similar pattern was illustrated where the
8
9 dispensations predicted to be higher risk had higher event rates (Figure 4).
10
11

12
13
14 After using the XGBoost and logistic regression classifiers to identify the dispensations in
15
16 the highest predicted risk percentiles, the pre-test probability of the outcome (1.6%) was
17
18 transformed into higher post-test probabilities, with higher probabilities in the riskier
19
20 percentiles (Table 1). The number needed to screen also decreased as predicted risk increased
21
22 (Table 1).
23
24

25
26 Comparing discrimination performance, ML risk prediction outperformed the current
27
28 guideline approaches when using various combinations of guideline recommendations (Table
29
30 2). In many of the guideline scenarios, the estimated AUROCs were close to the 0.5 mark.
31
32 When we estimated the discrimination performance of the logistic regression classifier based
33
34 on database source, using all databases produced an AUROC of 0.88. Reducing the database
35
36 source to only DAD, NACRS, Claims (co-morbidities only) resulted in an AUROC of 0.85, while
37
38 PIN (prescription history) only was 0.78 (Table 3).
39
40
41
42
43

44 Discussion

45
46
47 This study showed that ML techniques using available administrative data (prescription
48
49 histories and ICD codes) may provide enough discriminatory performance to predict adverse
50
51 outcomes associated with opioid prescribing. Indeed, our ML analyses showed very high
52
53 discrimination performance at 0.88. The linear model (logistic regression) and XGBoost carried
54
55
56
57

1
2
3 higher discrimination and calibration performance, while the neural network classifier did not
4
5 perform as well. By identifying the predicted top 5-10 percentile of absolute risk pursuant to an
6
7 opioid dispensation, we were able to capture approximately half of all outcomes using ML
8
9 methods. All ML models we trained had higher discrimination performance using the validation
10
11 sets compared to the clinical guideline approach.
12
13
14
15

16 Since the prevalence of our defined outcome is relatively low in the general population,
17
18 PPVs would also be expectedly low. However, estimated PPVs increased when we considered
19
20 higher risk dispensations, as is expected since PPV is related to event prevalence. This is
21
22 important because different users of a risk predictor will require different predictive
23
24 capabilities. Similarly, our estimates of positive likelihood ratios and associated post-test
25
26 probabilities also increased in dispensations with higher predicted risk indicating the strong
27
28 predictive capabilities of the XGBoost and logistic regression classifiers; likelihood ratios >10
29
30 generate conclusive changes from pre-test to post-test probabilities²⁰.
31
32
33
34
35

36 The current guideline approach to assess absolute opioid prescribing risk produced c-
37
38 statistic estimates closer to 0.5 indicating that discrimination was not much better than chance
39
40 alone. ML models with higher predictive performance can better support health departments
41
42 and PMPs with monitoring mandates to identify and intervene on those at high risk and their
43
44 associated prescribers. We also found that adding co-morbidity features from administrative
45
46 databases increased prediction performance compared to prescription history alone, thus
47
48 making the case for the use of this data by PMPs and health departments. However, if only
49
50 prescription history is available, our trained XGBoost classifier still had strong discrimination
51
52 performance.
53
54
55
56
57

1
2
3 We found only one study that used ML approaches to quantify the absolute risk of an
4 event pursuant to an opioid dispensation¹⁰. Their methodology used rolling 3-month windows
5 for estimating risk and ML model training while we used historic records to estimate 30-day
6 risk. Differences in study population and feature selection may explain why their highest
7 performing ML model was deep learning (neural network classifier) and ours was not.
8 Nevertheless, we were able to replicate their predictive performance using our ML approach as
9 we both showed that ML approaches have higher predictive capabilities than guideline
10 approaches. Both of our studies used predicted percentile risk estimates to identify high risk
11 dispensations and were able to do so with strong discrimination and calibration performance.
12 Furthermore, we emphasized prognostic metrics which are more informative to assess the
13 clinical utility of ML classifiers using pre- and post-test probabilities, something not done in
14 other studies and recommended in medical guidelines²⁰. This major aspect of our study, not
15 done previously, is important because any ML classifier that does not increase prognostic
16 information compared to baseline cannot be incorporated into decision making for the purpose
17 of intervening on higher risk instead of lower risk patients. Indeed, another study we found
18 describes how identifying cases in higher predicted risk percentiles using ML methods can be
19 deployed in hospital settings for the purpose of targeted interventions³⁹ upon discharge,
20 however the effect on outcomes is still to be determined.
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

48 The limitations of our study are similar to other ML studies¹⁰ and need to be addressed
49 when considering deployment of ML risk predictors. Our training dataset was not able to
50 account for non-prescription opioid consumption and the risk associated with non-prescription
51 use, both of which are substantial contributors to overall risk². Regarding our analysis, we
52
53
54
55
56
57

1
2
3 assumed that all dispensations were independent events; future research in this area should
4
5 focus on employing ML methods using correlated data. As with all ML projects, our models
6
7 were trained using Alberta data and might not be generalizable to other populations, or to
8
9 specific populations within Alberta. However, one of the benefits of the ML process is that
10
11 models can be retrained or similar methods could be used to develop new models to
12
13 accommodate different populations.
14
15
16
17

18 This study suggests that ML risk prediction can support PMPs, especially if readily
19
20 available administrative health data is used. PMPs currently use population-based guidelines
21
22 which we, and others, have shown cannot predict absolute individual risk. The ML process
23
24 allows for flexibility in model training, validation and deployment to specific settings in which,
25
26 for the case of PMPs, high risk patients can be identified and targeted for intervention either at
27
28 the patient or provider level. For example, a ML classifier can be trained on accessible data to
29
30 create an aggregated list of “high risk” patients at regular time intervals to identify points of
31
32 intervention. Moreover, ML classifiers can be retrained over time as changes in populations
33
34 and trends in prescribing occur and are therefore specific to the population unlike broadly
35
36 based guidelines. Further research can assess whether implementation of a ML-based
37
38 monitoring system by PMPs leads to improved clinical outcomes within their own jurisdictions
39
40 and whether other available features or feature reduction can yield sufficiently valid results for
41
42 their own intended purposes.
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 Pain Opioid 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. Hsieh 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 mining 2015.
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 systems 2017.
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 mining 2016.
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.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

*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

For peer review only

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 equivalents³⁶, 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

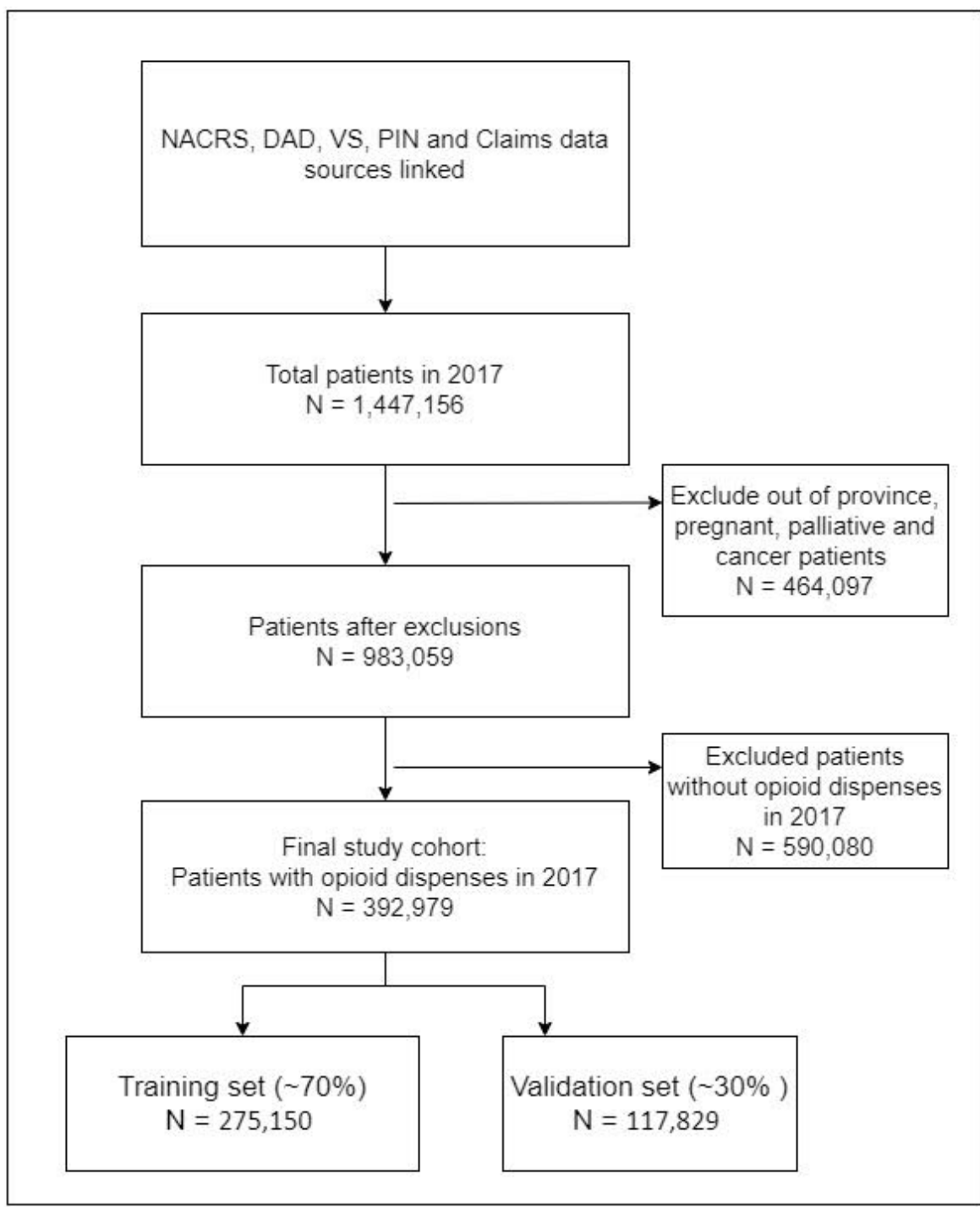
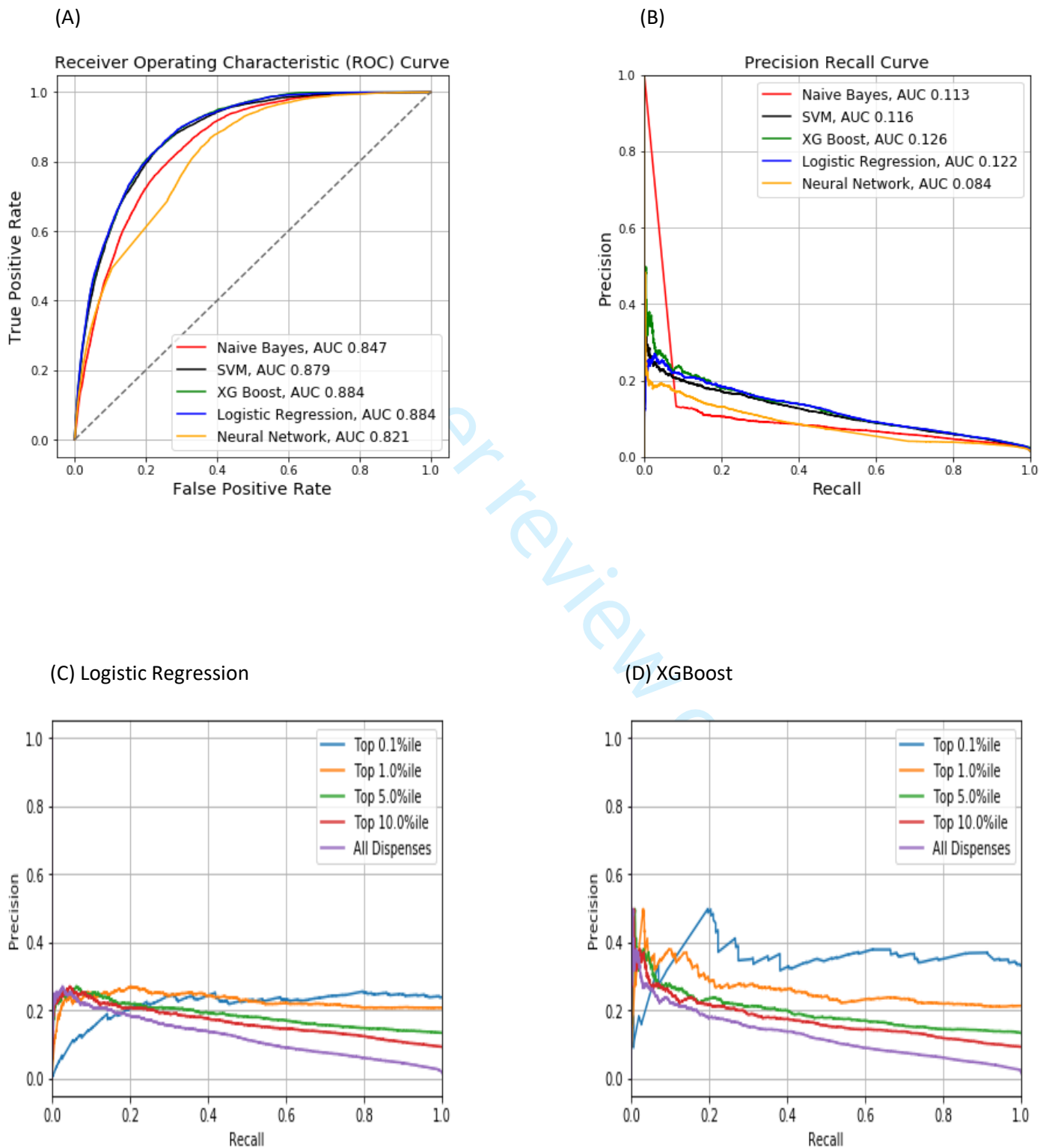


Figure 2. Area under the receiver operating characteristic curves (A) and precision-recall curves (B) for all dispensations using logistic regression (L1), neural network, support vector machine (SVM), XGBoost and Naïve-Bayes; precision-recall curves for higher risk dispensations according to predicted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



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.

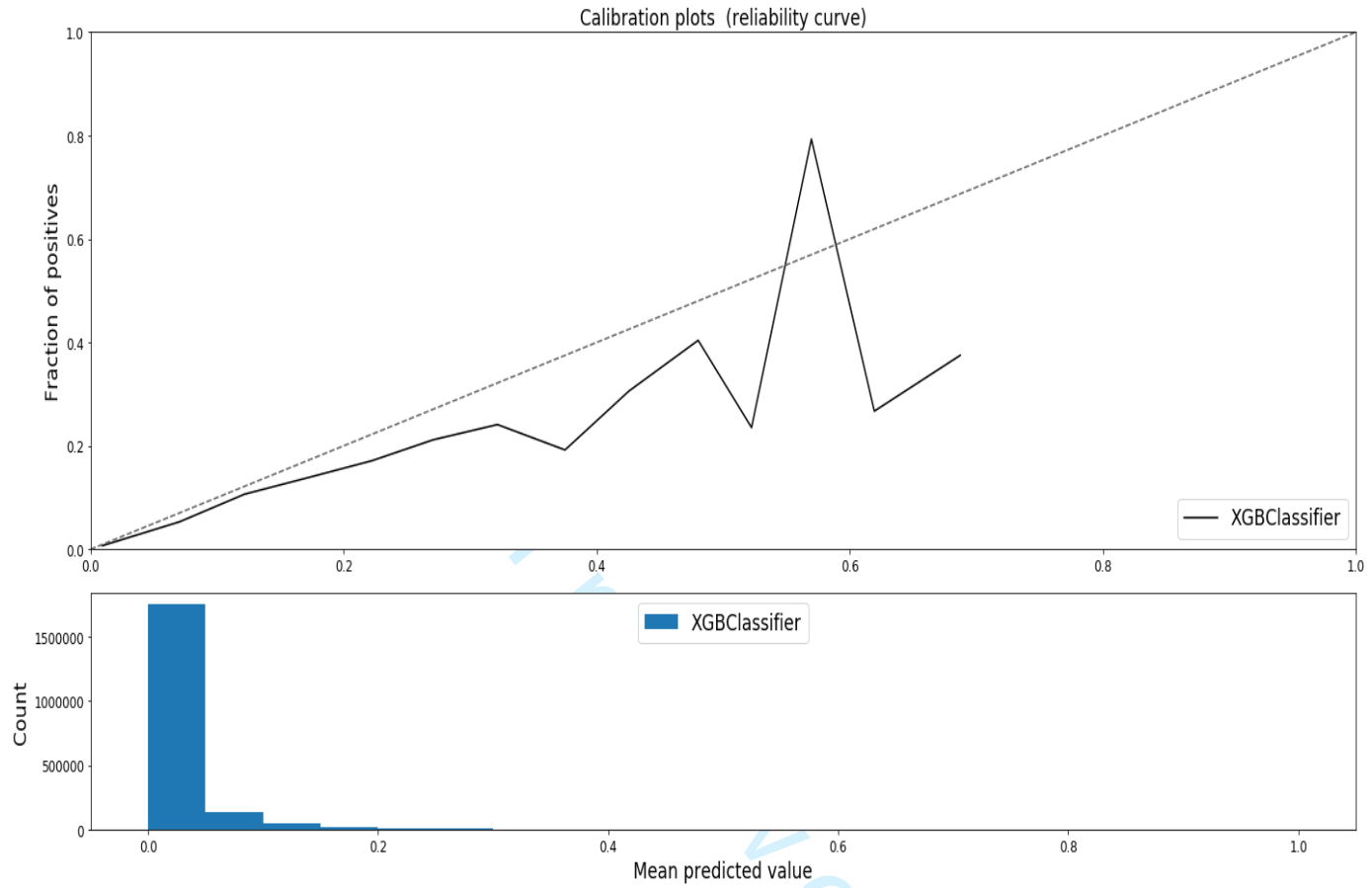
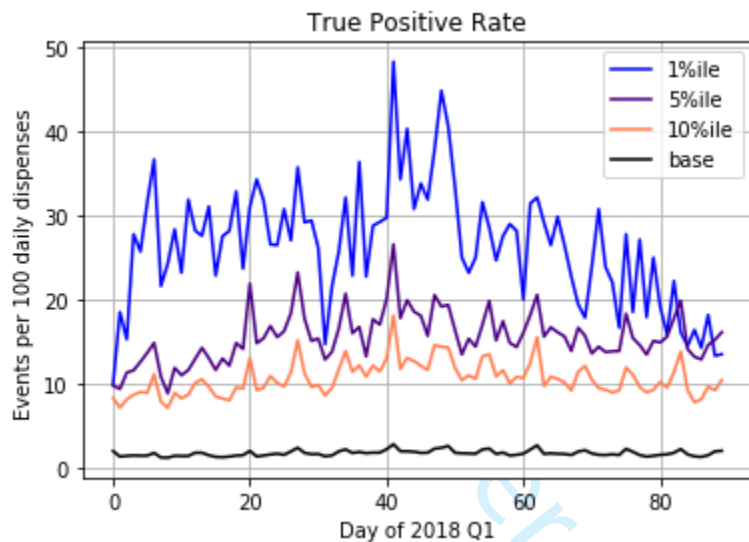
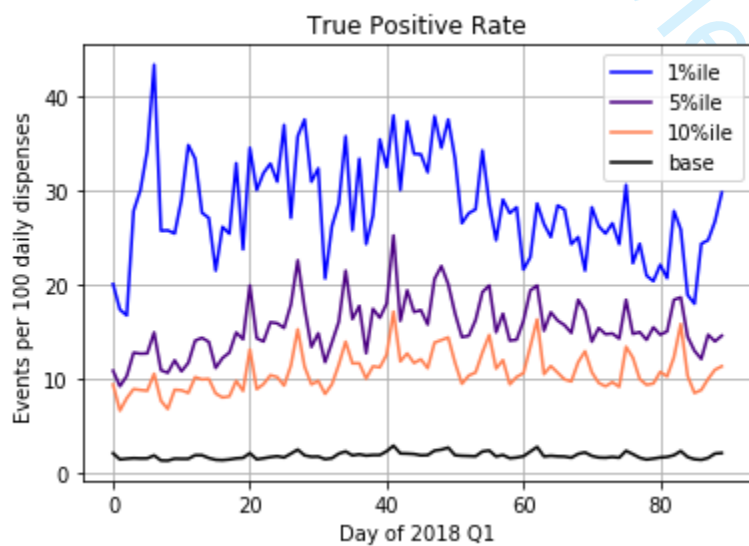


Figure 4. Simulation of a clinical workflow with daily uploads and events per 100 daily dispenses by risk percentiles using 2018 Quarter 1 (Q1) data for logistic regression (A) and XGBoost (B) classifiers.

(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 chi² 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

1
2
3 functions has shown to be very effective recently in many application domains such as natural
4 language processing, computer vision, genomics, computer games and health[2]. Neural
5 networks come in many flavors learning nonlinear mapping of different types of data such as
6 Convolutional NNs being most effective with images and Recurrent NNs for time series and
7 language data. Identifying the most effective neural network structure is one of the difficult and
8 the most time-consuming aspect of applying neural networks to new application domains and
9 data. Generally, neural networks try to exploit the relationships in the raw unstructured data (eg:
10 image and text) presented to the network but with more structured data such as health records
11 and ICD codes learning relationships is much complex. Our neural network models are mainly
12 based on densely connected hidden layers with ReLu[6] activation function. We used the cross-
13 entropy loss for the binary classification Adam optimizer. We used a simple feed forward
14 network using Sklearn MLP classifier with hyperparameter tuning for the NN.
15
16
17
18

19 **Boosted Learning Algorithms**

20
21 Boosting is a process to ensemble multiple base learning algorithms to arrive at better overall
22 performance than any individual base learner[1]. In contrast to independently building multiple
23 models from the subsets of the data, boosting re-weights the training data every time a model is
24 learned for future models. This weighting happens to give more preference to currently
25 misclassified data points in the next round compared to the correctly classified data points.
26 Therefore future learners try to do better on the misclassified data points leading to a collection
27 base learners having a better-combined prediction. This process is sequential so each base
28 learner is dependent on the output of the previously trained model (it is worthy to note XGBoost
29 provides a parallel tree boosting alternative). In our work, we have experimented with several
30 boosting meta-learning algorithms such as XGBoost[7], AdaBoost[5] and GBM[5]. XGBoost uses
31 a variant of trees as the base learner whereas AdaBoost (from Sci-kit learn) can use many ML
32 algorithms as base learners. GBM uses logistic regression by default as the base learner. We used
33 all 3 types of boosting with tuned hyperparameters for comparison.
34
35
36
37
38

39 **Naive Bayes**

40
41 Naive Bayes is based on the Bayes theorem with a strong independence assumption between the
42 covariates[1]. This assumption helps in building a simple probabilistic model for learning and
43 inference. Naive Bayes coefficients scale linearly with the number of covariates making this a
44 suitable model for high-dimensional data. We used Naive Bayes as a simple baseline learning
45 algorithm for comparison.
46
47
48

49 **Support Vector Machines (SVM)**

50
51 SVMs[4] are maximum margin classifiers optimizing for learning a hyperplane having the
52 maximum distance away from each of the class data points[1]. SVM is a linear classifier but with
53 the kernel trick to map the inputs to the higher dimensional space, it can learn nonlinear decision
54 boundaries in the input space. SVMs are very effective binary classifiers with the kernel trick[1].
55 With larger datasets, SVMs tend to become more computationally intensive.
56
57
58

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 χ^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

1
2
3 ^ Injury: ICD10: S00-T98; Poisoning: ICD10: T36-T50

4 ** Complicated, uncomplicated diabetes and hypertension were collapsed into one category each
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

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 χ^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.

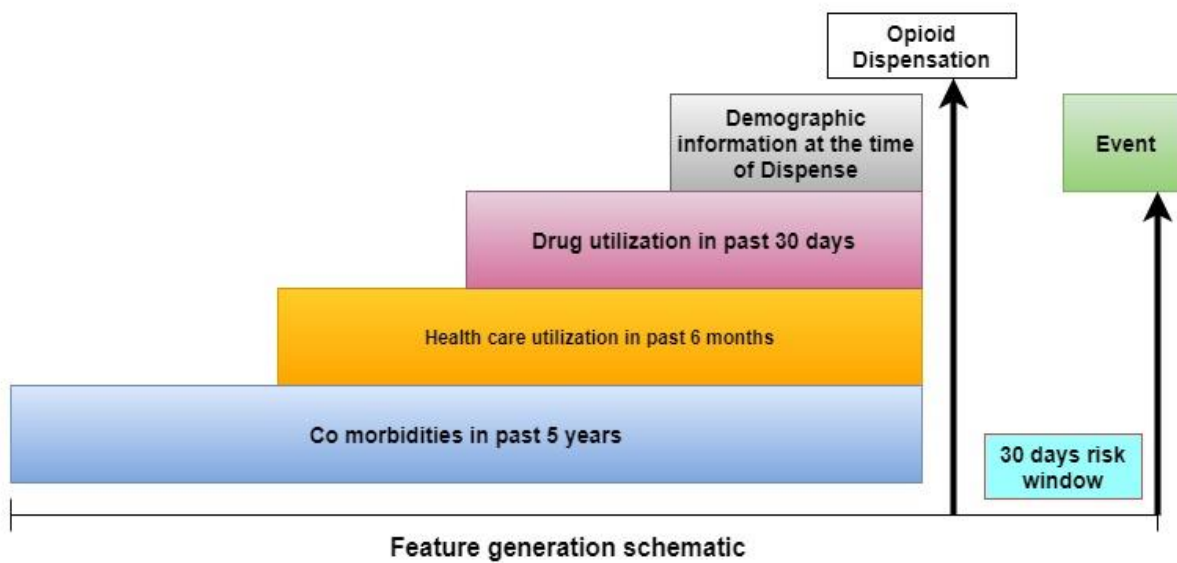
Note: ATC- Anatomical Therapeutic Chemical classification (https://www.whocc.no/atc_ddd_index); PIN- Pharmaceutical Information Network; ICD- International Statistical Classification of Diseases and Related Health Problems, World Health Organization; total number of features 283

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



review only

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

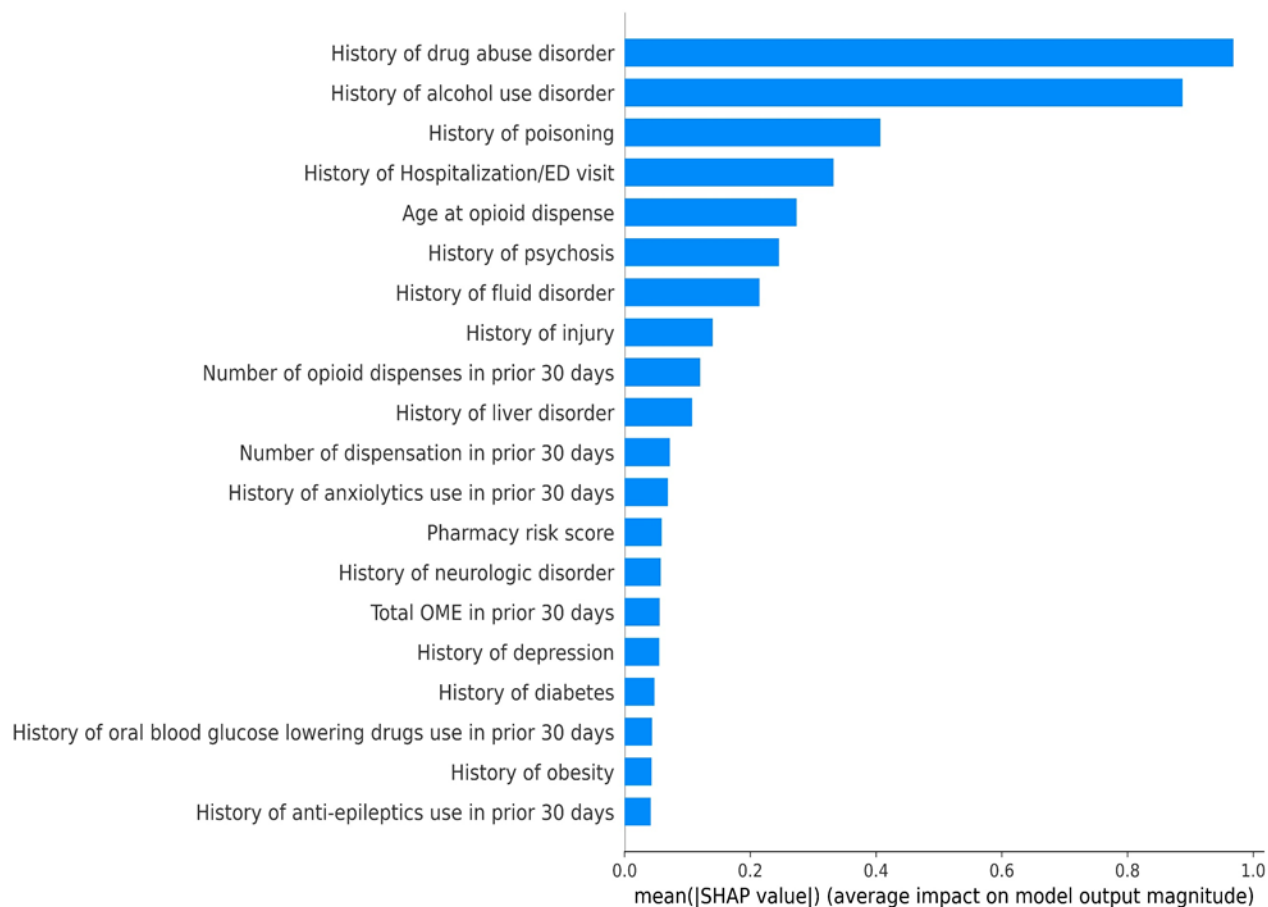
Note: Logistic regression used L1 (lasso) parameter regularization; BZD- benzodiazepine; OME- oral morphine equivalents; ED: emergency department

*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

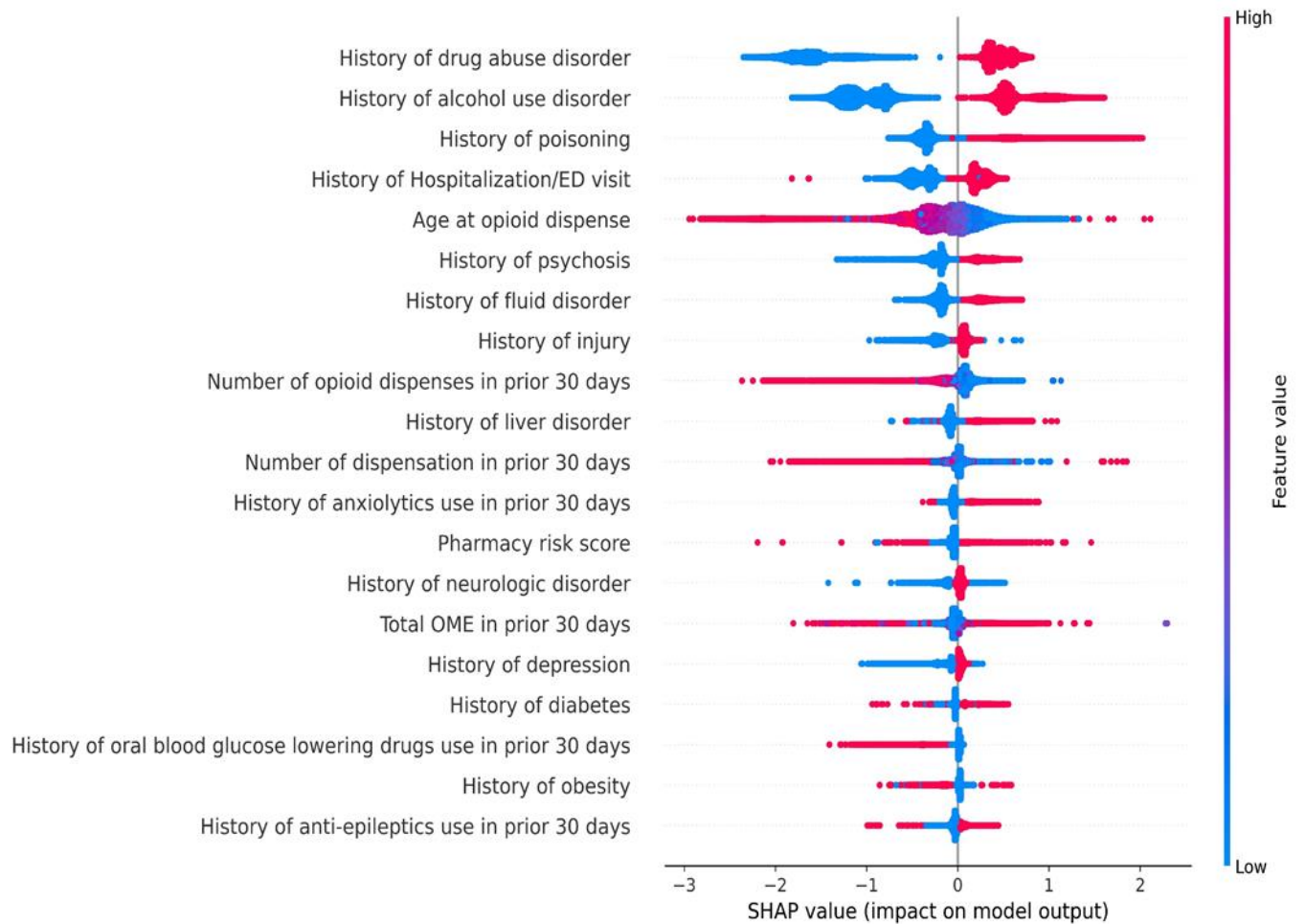
eFigure 3. SHAP values and feature impact of the XGBoost classifier using the 2018 validation set to describe “associations” between features and the outcome. Features with the most impact on the model with drug abuse ranked highest (A); tornado plot illustrating feature impact (B); explaining the prediction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

(A)



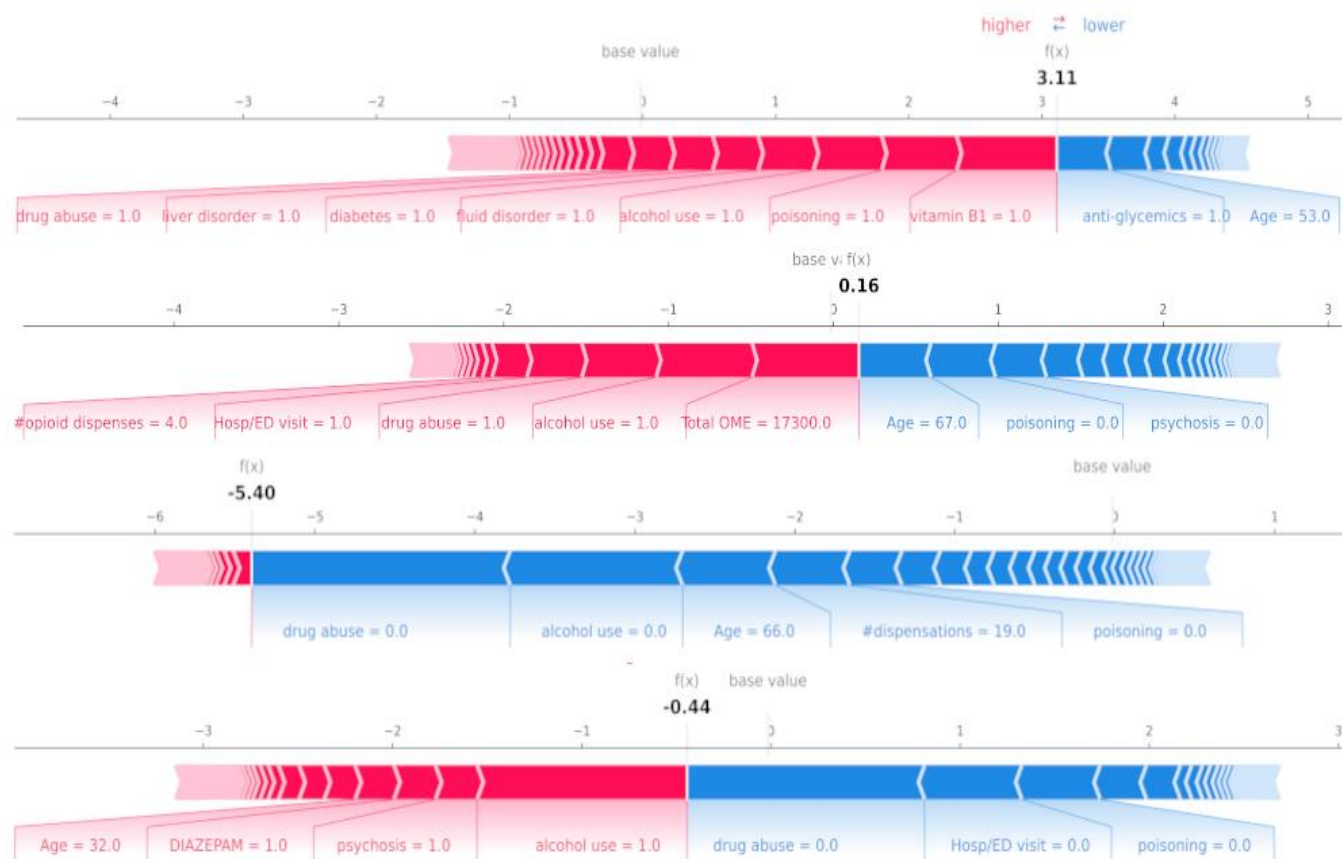
Note: Pharmacy risk score- derived feature using proportion of opioid patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each pharmacy; training and validating the XGBoost classifier with these features alone resulted in an AUC of 0.877 in the 2018 validation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

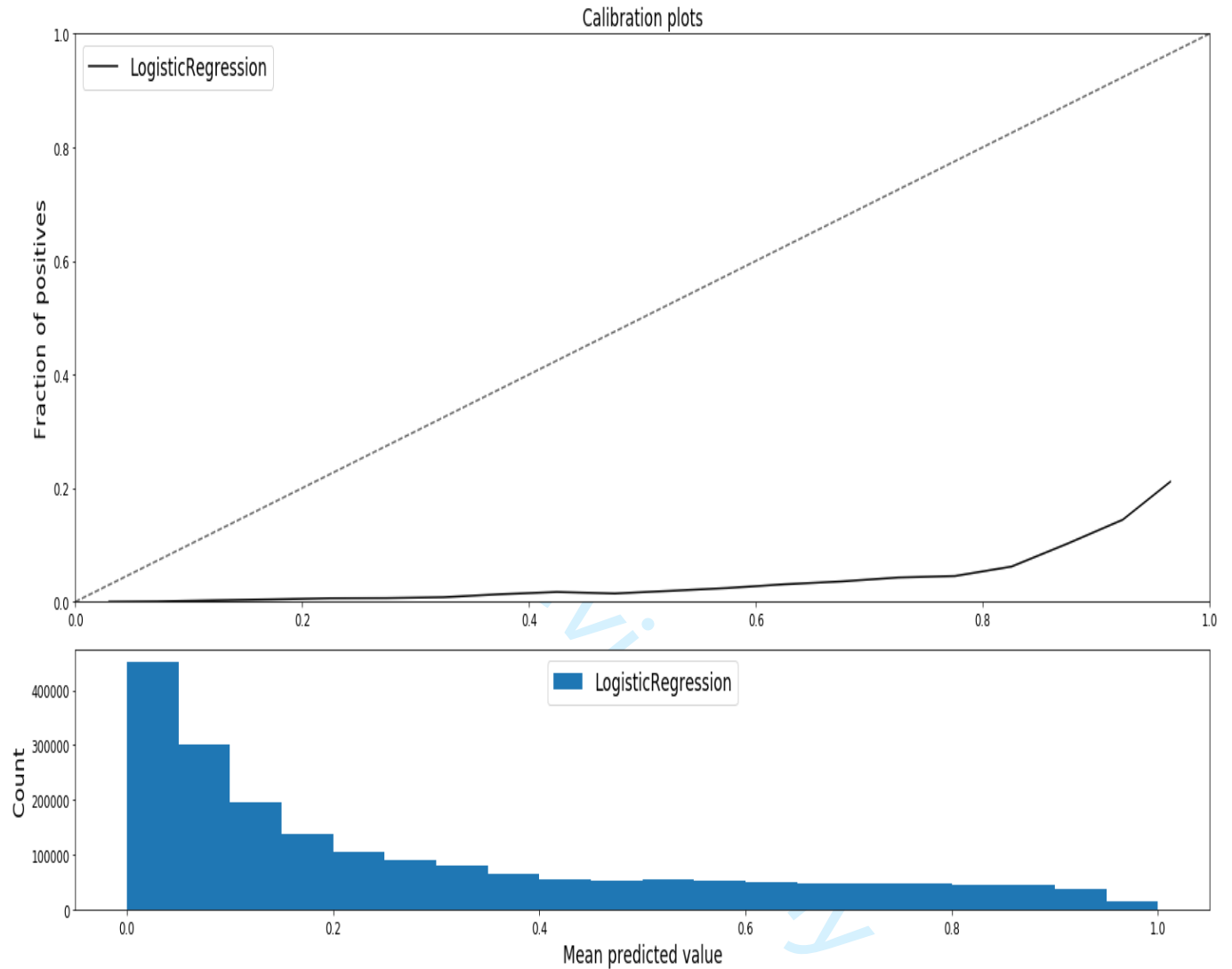


How to read this figure: Using hospitalization within 30-days of an opioid dispensation as the outcome of interest, there are 4 scenarios to consider: the XGBoost classifier has low or high confidence in predicting a hospitalization and low or high confidence in predicting **NO** hospitalization. Start 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

1
2
3 **eFigure 4.** Calibration curve plotting observed vs. quantiles of estimated risk for the logistic regression
4 (L1) classifier using the 2018 validation dataset. The majority of counts (dispensations) were predicted
5 to be lower risk.
6
7
8
9



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:	<i>BMJ Open</i>
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 Kulkarni, Vinaykumar; OKAKI Health Analytics Eurich, Dean; University of Alberta, School of Public Health Kumar, Luke; Alberta Machine Intelligence Institute 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

SCHOLARONE™
Manuscripts



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.

1
2
3 **Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk**
4 **after an opioid dispensation in Alberta, Canada**
5

6 **Author list (in order):**
7

8 Vishal Sharma (0000-0001-7907-1183), Vinaykumar Kulkarni, Dean T. Eurich (0000-0003-2197-
9 0463), Luke Kumar, Salim Samanani (0000-0001-6751-4805)
10
11
12

13
14 **Address for each author:**
15

16 2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of
17 Alberta, Edmonton, Alberta, Canada, T6G 2E1 Vishal Sharma BPharm PhD Candidate,
18
19

20
21 OKAKI Health Intelligence, Edmonton, Alberta, Canada, Vinaykumar Kulkarni MSc
22
23

24
25 2-040 Li Ka Shing Center for Health Research Innovation, School of Public Health, University of
26 Alberta, Edmonton, Alberta, Canada, T6G 2E1 Dean Eurich professor
27
28

29
30 Alberta Machine Intelligence Institute, Edmonton, Alberta, Canada, T5J 3B1 Luke Kumar MSc
31
32

33
34 OKAKI Health Intelligence, Calgary, Alberta, Canada, Salim Samanani MD, Medical Director
35
36
37

38 **Corresponding Author:**
39

40 Dean Eurich, 2-040 Li Ka Shing Center for Health Research Innovation, University of Alberta,
41 Edmonton, Alberta, Canada, T6G 2E1; Phone 780-492-6333; fax 780-492-7455; email:
42 deurich@ualberta.ca
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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.

Copyright/license for publication: *The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide licence to the Publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) licence any third party to do any or all of the above.*

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).

1
2
3 **Data Sharing:** The data used in this study is not available for external analysis. However, administrative
4 health data can be accessed from Alberta Health by following defined research protocols and
5 confidentiality agreements.
6
7

8
9 **Transparency:** The lead author, VS, (the manuscript's guarantor, Dean Eurich) affirms that the
10 manuscript is an honest, accurate, and transparent account of the study being reported; that no
11 important aspects of the study have been omitted; and that any discrepancies from the study as
12 originally planned (and, if relevant, registered) have been explained.
13
14

15
16 **Word Count: 3357**
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

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 crisis¹; a major part of the policy response to the opioid crisis focuses on endorsing safe, appropriate opioid prescribing²⁻⁴. 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 populations¹¹. 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.

1
2
3 Supervised machine learning (ML)^{12,13} is an approach that uses computer algorithms to
4
5 build predictive models in the clinical setting that can make use of the large amounts of
6
7 available administrative data^{14,15}, all within a well-defined process¹⁶. Supervised ML trains on
8
9 labelled data to develop prediction models that are specific to different populations and, in
10
11 many cases, can provide better predictive performance than traditional, population-based
12
13 statistical models^{10,15,17}. We identified one study¹⁰ that applied ML techniques to predict
14
15 overdose risk in opioid patients pursuant to a prescription. In their validation sample, they
16
17 found that the deep neural network (DNN) and gradient boosting machines (GBM) algorithms
18
19 carried the best discrimination performance based on estimated c-statistics and that the ML
20
21 approach out-performed the guideline approach in terms of risk prediction; neural networks
22
23 have little interpretability and are not necessarily better at predicting outcomes when trained
24
25 on structured data¹⁸. This study relied on c-statistics to evaluate their ML models and did not
26
27 emphasize other performance metrics (e.g., positive likelihood ratios, pre and post-test
28
29 probabilities) required to assess clinical utility that are recommended by medical reporting
30
31 guidelines^{11,13,19,20}. It also did not address the important issue of ML model interpretability²¹.
32
33 Reporting informative prognostic metrics is needed to better understand the capabilities of ML
34
35 classifiers if health departments and PMPs are to incorporate them into their decision-making
36
37 processes.
38
39
40
41
42
43
44
45
46
47

48 The objective of our study was to further develop and validate ML algorithms (beyond
49
50 just DNN) to predict the 30-day risk of hospitalization, emergency visit and mortality for a
51
52 patient in Alberta, Canada at the time of an opioid dispensation using administrative data
53
54 routinely available to health departments and PMPs and evaluate them using the above
55
56
57

1
2
3 referenced reporting guidelines. We also analyzed feature importance to provide meaningful
4
5 interpretations of the ML models. Comparing discrimination performance (area under the
6
7 receiver operating characteristics curves), we hypothesized that the ML process would perform
8
9 better than the current guideline approach for predicting risk of adverse outcomes related to
10
11 opioid prescribing.
12
13
14
15

16 **Methods**

17 **Study Design and Participants**

18
19
20
21
22 This prognostic study used a supervised ML scheme. All patients in Alberta, Canada who
23
24 received a dispensation for an opioid, were 18 years of age and older between Jan 1, 2017 and
25
26 Dec 31, 2018 were eligible. Patients were excluded from all analyses if they had any previous
27
28 diagnosis of cancer, received palliative interventions or were pregnant during the study period
29
30 (eTable 1 in Supplement) as use of opioids in these contexts is clinically different.
31
32
33
34

35
36 Government health departments and payers in many jurisdictions have systems to
37
38 capture prescription histories and ICD diagnostic codes. As such, we linked various
39
40 administrative health data sets available in Alberta, Canada using unique patient identifiers in
41
42 order to establish a complete description of patient demographics, drug exposures and health
43
44 outcomes. These databases include 1) *Pharmaceutical Information Network (PIN)*: PIN data
45
46 includes all dispensing records from community pharmacies from all prescriber types occurring
47
48 in the province outside of the hospital setting. PIN collects all drug dispensations irrespective of
49
50 age or insurance status in Alberta; Anatomical Therapeutic Chemical classification (ATC) codes²²
51
52 were used to identify opioid dispensations and their respective opioid molecules (eTable 5), 2)
53
54
55
56
57
58
59
60

1
2
3 *Population and Vital Statistics Data (VS, Alberta Services):* sex, age, date of birth, death date,
4
5 immigration and emigration data, and underlying cause of death according to the World Health
6
7 Organization algorithm using ICD codes⁸, 3) *Hospitalizations and Emergency Department Visits*
8
9 *(National Ambulatory Care Reporting System [NACRS], Discharge Abstract Database [DAD]):* all
10
11 services, length of stay, diagnosis (up to 25 ICD-10⁸ based diagnoses). Data and coding accuracy
12
13 are routinely validated both provincially and centrally via the Canadian Institute for Health
14
15 Information, and 4) *Physician Visits/Claims (Alberta Health):* all claims from all settings (e.g.,
16
17 outpatient, office visits, emergency departments, inpatient) with associated date of service, ICD
18
19 code, procedure and billing information.
20
21
22
23
24
25

26 This study followed the TRIPOD and STARD reporting guidelines²³⁻²⁵ and received ethics
27
28 approval from the University of Alberta ethics board (Pro00083807_AME1).
29
30

31 **Measures and Outcome**

32
33

34 ML models were trained on a labelled dataset in which the observation/analysis unit
35
36 was an opioid dispensation. Every opioid dispensation, not just the incident one, was used as a
37
38 potential instance to predict the risk of our outcome. The primary outcome was a composite of
39
40 a drug-related hospitalization, emergency department (ED) visit or mortality within 30 days of
41
42 an opioid dispensation based on ICD-10 codes used by others and identified from DAD, NACRS
43
44 and Vital Statistics (T40, F55, F10-19; eTable 2 in Supplement)^{2,10,26}.
45
46
47
48
49

50 We anticipated that our defined outcome would be a rare event, leading to a class
51
52 imbalanced dataset²⁷. To address this, we relied on specifying balanced class weightage for
53
54 supporting algorithms; other approaches were deemed not suitable (e.g., oversampling using
55
56
57

1
2
3 randomly repeating minority class); under sampling (sub-sampling within the majority class)
4
5 resulted in changes in outcome prevalence. Class weightage is a commonly used method²⁸ to
6
7 address class imbalance along with over and under-sampling approaches. However,
8
9
10 oversampling, which involves generating new opioid dispensations from the original data
11
12 distribution and is prone to introducing bias, is difficult due to the categorical nature of the data
13
14 and beyond the scope of this study. With under-sampling, which takes samples from the
15
16 majority class (in this case, no 30-day event after dispensation), we would not be able to use all
17
18 of the information provided by the data in instances with no outcome. Hence, we decided to
19
20 use the class weightage method which does not alter the data distribution. Instead, the
21
22 learning process is adjusted in a way that increases the importance of the positive class
23
24 (instances that led to a 30-day event)²⁹.
25
26
27
28
29
30

31 **Predictor Candidates for ML Models**

32
33
34 Predictor variables in our ML models included those that were informed by the
35
36 literature^{3,4,10} and those directly obtained from the data sets. These included features based on
37
38 demographics (age, sex, income using Forward Sortation index from postal codes³⁰), co-
39
40 morbidity history using ICD-based Elixhauser score categories³¹, health care utilization (number
41
42 of unique providers, number of hospital and emergency department visits), and drug utilization
43
44 (level 3 ATC codes²², oral morphine equivalents³², concurrent use with benzodiazepines,
45
46 number of opioid and benzodiazepine dispensations, number of unique opioid and
47
48 benzodiazepine molecules). Depending on the potential predictor and data availability, we
49
50 used data from 30 days to 5 years before the opioid dispensation to generate model features
51
52 (eFigure 1 in Supplement); 30 days was used to reflect the immediate nature of the risk and 5
53
54
55
56
57
58
59
60

1
2
3 years to fully capture co-morbidities. This approach aligns with how health providers would
4
5 assess patients using the entire history of co-morbidities and then the more immediate factors
6
7 in deciding on the need for a therapeutic as well as risk in patients. We performed experiments
8
9 to identify the features and data sets that contributed most to predicting the outcomes with a
10
11 view to minimizing the potential future data requirements for health departments and PMPs.
12
13
14
15

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

17
18
19 We randomly divided the patients in the 2017 portion of our study cohort into training
20
21 (70%) and validation (30%) sets¹³ by patients and opioid dispensations such that no patients in
22
23 the training set were in the validation set. Baseline characteristics and event rates were
24
25 compared in the training vs validation group, and between those who experienced the outcome
26
27 and those who did not using chi-squared tests of independence. As well, we used all the 2018
28
29 data as another independent validation set.
30
31
32
33

34 We trained commonly used^{13,33} ML algorithms (eAppendix in Supplement) and further
35
36 tuned out-of-box models using 5-fold cross validation on the training data to address model
37
38 overfitting^{13,34}. As is common in ML validation studies^{10,13}, we reported model discrimination
39
40 performance (i.e. how well a model differentiates those at higher risk from those at lower
41
42 risk)¹¹ using area under the receiver operating characteristic curve (AUROC; c-statistic). We
43
44 then stratified the two ML models with the highest c-statistics into percentile categories
45
46 (deciles) according to absolute risk of our outcome, as was done in previous studies^{10,35}. We
47
48 also plotted AUROC¹¹ and precision-recall curves (PRCs)³⁶.
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Because discrimination alone is insufficient to assess ML model prediction capability, we
4
5 assessed a second necessary property, namely, calibration (i.e., how similar the predicted
6
7 absolute risk is to the observed risk across different risk strata)^{11,37}. Using the two ML models
8
9 with the highest discrimination performance, we assessed calibration performance on the 2018
10
11 data by plotting observed (fraction of positives) vs predicted risk (mean predicted value). Using
12
13 these same two ML classifiers, we analyzed the top 0.1, 1, 5, and 10 percentiles of predicted
14
15 risk by the number of true and false positives, positive likelihood ratios (PLR)²⁰, positive
16
17 predictive values (PPV), post-test probabilities, and number needed to screen. We also
18
19 performed a simulation of daily data uploads for 2018 Quarter 1 to view the predictive
20
21 capabilities if a ML risk predictor were to be deployed into a monitoring workflow.
22
23
24
25
26
27

28 For the XGBoost and logistic regression classifiers, we reported feature importance³³
29
30 and plotted PRCs that compared all dispenses to those within the top 10 percentiles of
31
32 estimated risk. As well, for the XGBoost classifier, we described feature importance on model
33
34 outcome using SHAP values^{38,39} to add an additional layer of interpretability.
35
36
37

38 Finally, we compared ML risk prediction (the two ML models with highest discrimination
39
40 performance) to current guideline approaches as others have¹⁰, using the 2019 Centers for
41
42 Medicare & Medicaid Services (CMS) opioid safety measures⁴⁰ and the 2017 Canadian Opioid
43
44 Prescribing Guideline³. This was done by using the guidelines as “rules” when coding for the
45
46 30-day risk of event at the time of each opioid dispensation on the entire 2018 validation set.
47
48 We also compared the discrimination performance of different logistic regression classifier
49
50 models using various combinations of features derived from their respective databases: **1)**
51
52
53
54
55
56
57

1
2
3 demographic and drug/health utilization features from PIN and 2) co-morbidity features
4
5 derived from DAD, NACRS and Claims.
6
7

8
9 All analyses were done using Python (v. 3.6.8,), SciKit Learn⁴¹ (v. 0.23.2) SHAP³⁹ (v. 0.35),
10
11 XGBoost (v. 0.90)⁴², Pandas (v. 1.0.5)⁴³ and H2O Driverless AI (version 1.9).
12
13

14 **Patient and Public Involvement**

15
16
17 This research was done without patient involvement. Patients were not invited to
18
19 comment on the study design and were not consulted to develop patient relevant outcomes or
20
21 interpret the results. Patients were not invited to contribute to the writing or editing of this
22
23 document for readability or accuracy. There are no plans to disseminate the results of the
24
25 research to study participants.
26
27
28
29

30 **Results**

31 32 33 **Patient Characteristics and Predictors**

34
35
36 We identified 392,979 patients with at least one opioid dispensation in 2017 (Figure 1).
37
38 This cohort was used to train (n= 275,150, 70%) and validate (n=117,829, 30%) ML models. In
39
40 2017 and 2018, 6,608 and 5,423 patients experienced the defined outcome, respectively.
41
42
43 Baseline characteristics were different between those who experienced the outcome and those
44
45 who did not (eTable 3 in Supplement) while characteristics were similar between the training
46
47 and validation sets (eTable 4 in Supplement). There were 2,283,075 opioid dispensations in
48
49 2017 and 1,977,389 in 2018. Overall, in 2017, 2.03% (n= 45,757) of opioid dispensations were
50
51 associated with the outcome; in 2018, the estimate was 1.6% (n= 31,392).
52
53
54
55
56
57
58
59
60

1
2
3 As described above, we categorized our candidate features into four groups (eTable 5 in
4 Supplement). When using all of the databases, the total number of features was 283 and 34
5
6 when considering only co-morbidities.
7
8
9

10 11 **Machine-Learning Prediction Performance** 12

13
14 Using the 2017 validation set, AUROCs for the XGBoost and logistic regression classifiers
15 had the highest discrimination performance at 0.87, while the neural network classifier had
16 lower performance at 0.80 (eTable 6 in Supplement).
17
18
19
20
21

22 Discrimination performance was similar for the 2018 validation set (n=393,023; eTable 6
23 in Supplement). XGBoost and logistic regression had the highest estimated AUROCs and area
24 under PRCs while the neural network classifier was lower (Figure 2A, 2B). As expected,
25 precision-recall curves indicate stronger predictive performance in opioid dispensations at
26 higher predicted risk percentiles (Figure 2C, 2D).
27
28
29
30
31
32
33
34

35 In the 2018 validation set, although discrimination performance was similar (0.88),
36 individual feature importance was different between the logistic regression and XGBoost
37 classifiers, with logistic regression feature importance more reliant on co-morbidity data from
38 DAD, NACRS and Claims while XGBoost relied more on drug utilization data from PIN (eFigure
39 2). With the XGBoost classifier, history of drug abuse, alcoholism, and prior
40 hospitalization/emergency visit carried the highest importance for predicting the study
41 outcome (eFigure 3A) where the presence of these features in a patient suggested a strong
42 prediction towards having the defined outcome (eFigure 3B and 3C).
43
44
45
46
47
48
49
50
51
52
53
54

55 **Calibration** 56 57

1
2
3 When considering dispensations predicted to be in the highest percentiles of risk, the
4 top 5-percentile captured 42% of all outcomes using the XGBoost and logistic regression
5 classifiers (Table 1). Also, as the predicted risk percentiles get higher (top 10 percentile to top
6 0.1 percentile), so too do the corresponding PPVs with the top 0.1 percentile associated with a
7 PPV of 33% for the XGBoost classifier. As well, lower categories of risk percentiles were
8 associated with lower outcomes (Figure 3, eFigure 4). When we simulated a monitoring
9 workflow scenario with daily data uploads, a similar pattern was illustrated where the
10 dispensations predicted to be higher risk had higher event rates (Figure 4).
11
12
13
14
15
16
17
18
19
20
21
22

23 After using the XGBoost and logistic regression classifiers to identify the dispensations in
24 the highest predicted risk percentiles, the pre-test probability of the outcome (1.6%) was
25 transformed into higher post-test probabilities, with higher probabilities in the riskier
26 percentiles (Table 1). The number needed to screen also decreased as predicted risk increased
27 (Table 1).
28
29
30
31
32
33
34
35

36 Comparing discrimination performance, ML risk prediction outperformed the current
37 guideline approaches when using various combinations of guideline recommendations (Table
38 2). In many of the guideline scenarios, the estimated AUROCs were close to the 0.5 mark.
39 When we estimated the discrimination performance of the logistic regression classifier based
40 on database source, using all databases produced an AUROC of 0.88. Reducing the database
41 source to only DAD, NACRS, Claims (co-morbidities only) resulted in an AUROC of 0.85, while
42 PIN (prescription history) only was 0.78 (Table 3).
43
44
45
46
47
48
49
50
51
52
53

54 Discussion

55
56
57

1
2
3 This study showed that ML techniques using available administrative data (prescription
4 histories and ICD codes) may provide enough discriminatory performance to predict adverse
5 outcomes associated with opioid prescribing. Indeed, our ML analyses showed very high
6 discrimination performance at 0.88. The linear model (logistic regression) and XGBoost carried
7 higher discrimination and calibration performance, while the neural network classifier did not
8 perform as well. By identifying the predicted top 5-10 percentile of absolute risk pursuant to an
9 opioid dispensation, we were able to capture approximately half of all outcomes using ML
10 methods. All ML models we trained had higher discrimination performance using the validation
11 sets compared to the clinical guideline approach.
12
13
14
15
16
17
18
19
20
21
22
23
24
25

26 Since the prevalence of our defined outcome is relatively low in the general population,
27 PPVs would also be expectedly low. However, estimated PPVs increased when we considered
28 higher risk dispensations, as is expected since PPV is related to event prevalence. This is
29 important because different users of a risk predictor will require different predictive
30 capabilities. Similarly, our estimates of positive likelihood ratios and associated post-test
31 probabilities also increased in dispensations with higher predicted risk indicating the strong
32 predictive capabilities of the XGBoost and logistic regression classifiers; likelihood ratios >10
33 generate conclusive changes from pre-test to post-test probabilities²⁰.
34
35
36
37
38
39
40
41
42
43
44
45

46 The current guideline approach to assess absolute opioid prescribing risk produced c-
47 statistic estimates closer to 0.5 indicating that discrimination was not much better than chance
48 alone. ML models with higher predictive performance can better support health departments
49 and PMPs with monitoring mandates to identify and intervene on those at high risk and their
50 associated prescribers. We also found that adding co-morbidity features from administrative
51
52
53
54
55
56
57

1
2
3 databases increased prediction performance compared to prescription history alone, thus
4
5 making the case for the use of this data by PMPs and health departments. However, if only
6
7 prescription history is available, our trained XGBoost classifier still had strong discrimination
8
9 performance.
10
11

12
13 We found only one study that used ML approaches to quantify the absolute risk of an
14
15 event pursuant to an opioid dispensation¹⁰. Their methodology used rolling 3-month windows
16
17 for estimating risk and ML model training while we used historic records to estimate 30-day
18
19 risk. Differences in study population and feature selection may explain why their highest
20
21 performing ML model was deep learning (neural network classifier) and ours was not.
22
23 Nevertheless, we were able to replicate their predictive performance using our ML approach as
24
25 we both showed that ML approaches have higher predictive capabilities than guideline
26
27 approaches. Both of our studies used predicted percentile risk estimates to identify high risk
28
29 dispensations and were able to do so with strong discrimination and calibration performance.
30
31 Furthermore, we emphasized prognostic metrics which are more informative to assess the
32
33 clinical utility of ML classifiers using pre- and post-test probabilities, something not done in
34
35 other studies and recommended in medical guidelines²⁰. This major aspect of our study, not
36
37 done previously, is important because any ML classifier that does not increase prognostic
38
39 information compared to baseline cannot be incorporated into decision making for the purpose
40
41 of intervening on higher risk instead of lower risk patients. Indeed, another study we found
42
43 describes how identifying cases in higher predicted risk percentiles using ML methods can be
44
45 deployed in hospital settings for the purpose of targeted interventions³⁵ upon discharge,
46
47 however the effect on outcomes is still to be determined.
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 The limitations of our study are similar to other ML studies¹⁰ and need to be addressed
4
5 when considering deployment of ML risk predictors. Our training dataset was not able to
6
7 account for non-prescription opioid consumption and the risk associated with non-prescription
8
9 use, both of which are substantial contributors to overall risk². Regarding our analysis, we
10
11 assumed that all dispensations were independent events; future research in this area should
12
13 focus on employing ML methods using correlated data. As with all ML projects, our models
14
15 were trained using Alberta data and might not be generalizable to other populations, or to
16
17 specific populations within Alberta. However, one of the benefits of the ML process is that
18
19 models can be retrained or similar methods could be used to develop new models to
20
21 accommodate different populations.
22
23
24
25
26
27

28 This study suggests that ML risk prediction can support PMPs, especially if readily
29
30 available administrative health data is used. PMPs currently use population-based guidelines
31
32 which we, and others, have shown cannot predict absolute individual risk. The ML process
33
34 allows for flexibility in model training, validation and deployment to specific settings in which,
35
36 for the case of PMPs, high risk patients can be identified and targeted for intervention either at
37
38 the patient or provider level. For example, a ML classifier can be trained on accessible data to
39
40 create an aggregated list of “high risk” patients at regular time intervals to identify points of
41
42 intervention. Moreover, ML classifiers can be retrained over time as changes in populations
43
44 and trends in prescribing occur and are therefore specific to the population unlike broadly
45
46 based guidelines. Further research can assess whether implementation of a ML-based
47
48 monitoring system by PMPs leads to improved clinical outcomes within their own jurisdictions
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 and whether other available features or feature reduction can yield sufficiently valid results for
4
5
6 their own intended purposes.
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

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 Pain Opioid 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 mining 2015.
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 systems 2017.
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 mining 2016.
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

1
2
3 OME: daily oral morphine equivalents; BZRA: benzodiazepine receptor agonist. Elixhauser scoring ICD
4 codes were used to identify mental disorders and substance abuse.

5 *The Canadian guidelines do not specify timelines. >90 OME was determined by taking the average
6 daily OME over the 30 days prior to dispensation
7

8 **The CMS guidelines specify 90 or more days at >120 OME and concurrent use of opioids and
9 benzodiazepines for 30 days or more within an assessment period of 180 days.
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57

For peer review only

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 equivalents³², 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

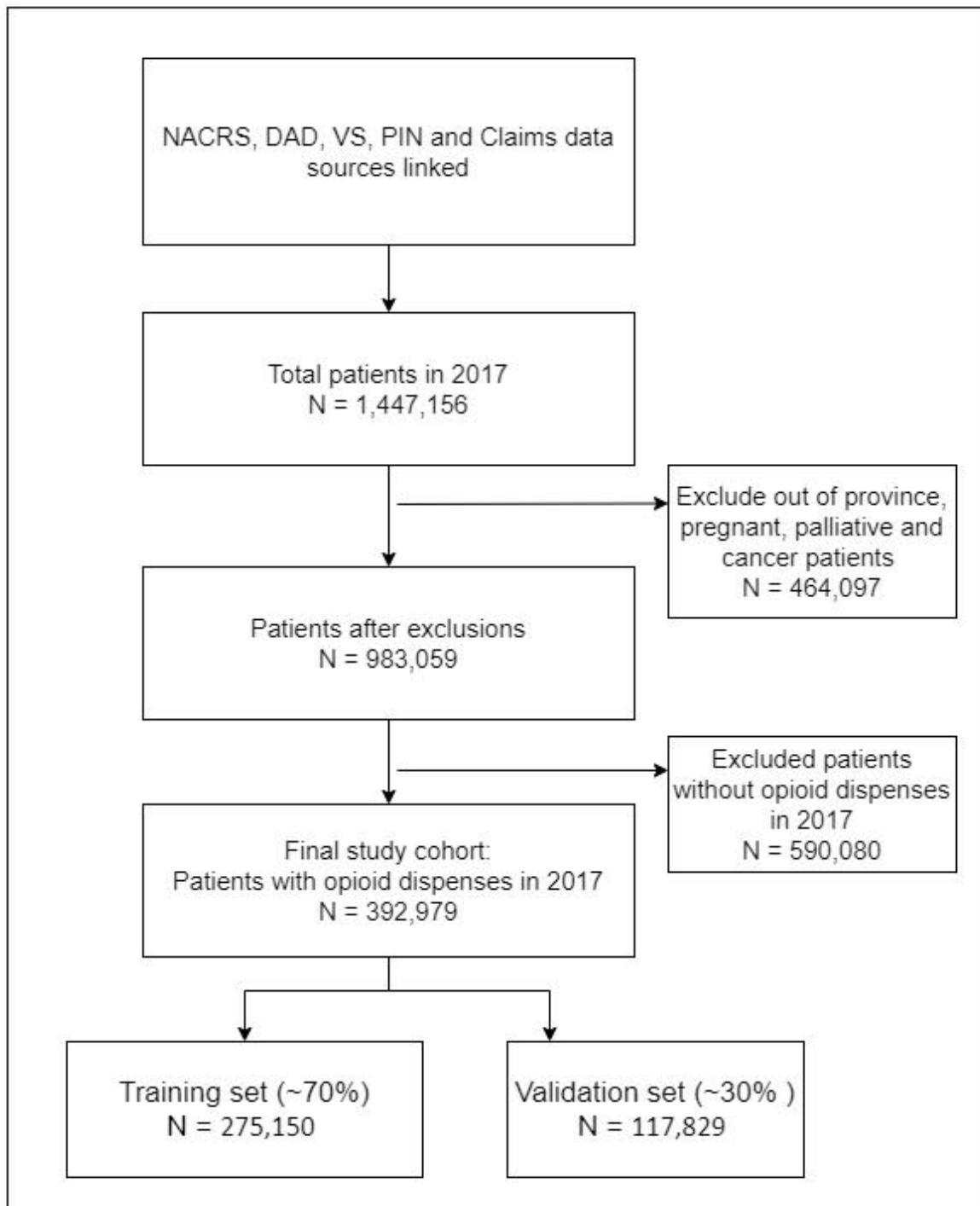
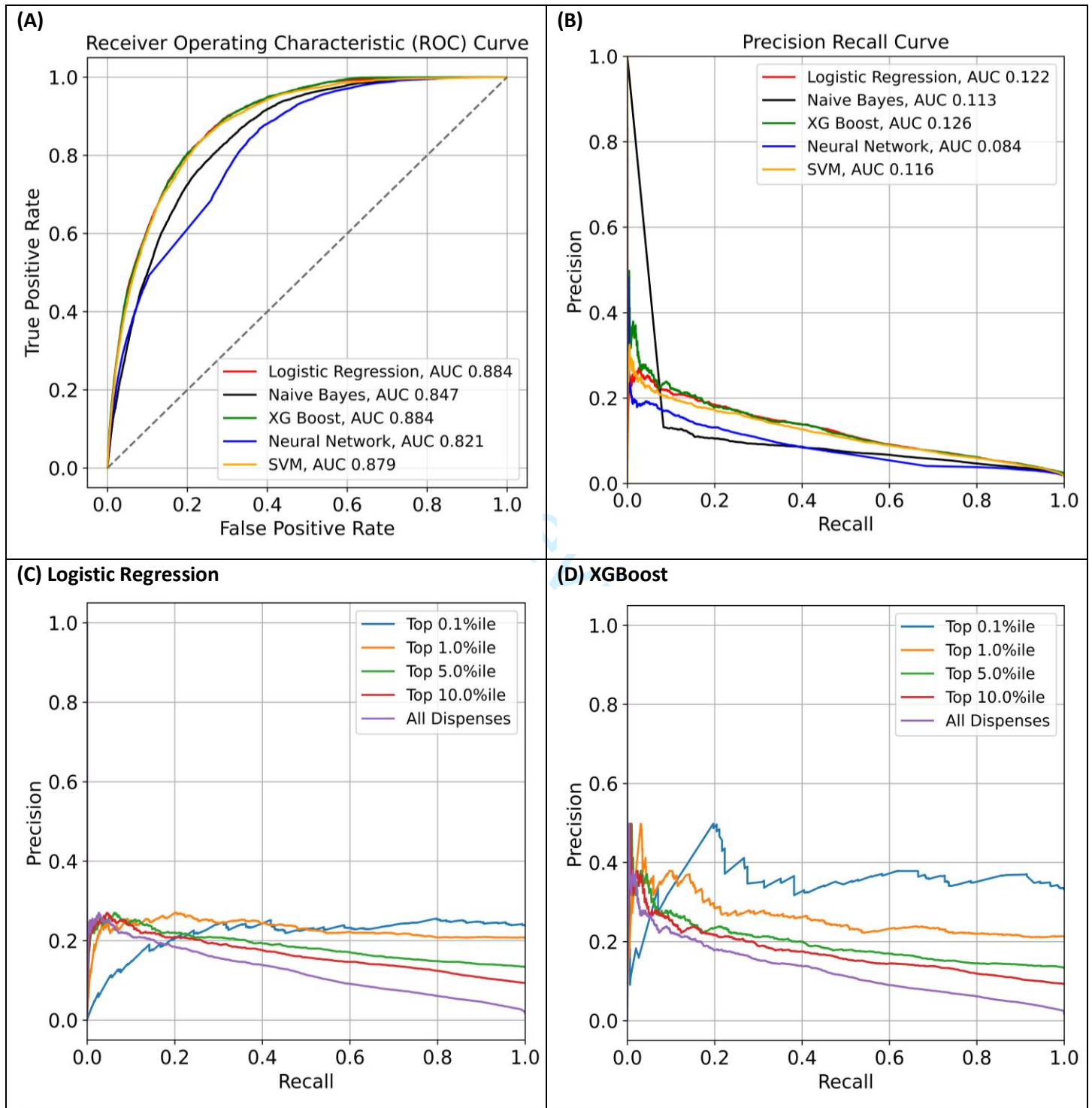


Figure 2. Area under the receiver operating characteristic curves (A) and precision-recall curves (B) for all dispensations using logistic regression (L1), neural network, support vector machine (SVM), XGBoost and Naïve-Bayes; precision-recall curves for higher risk dispensations according to predicted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

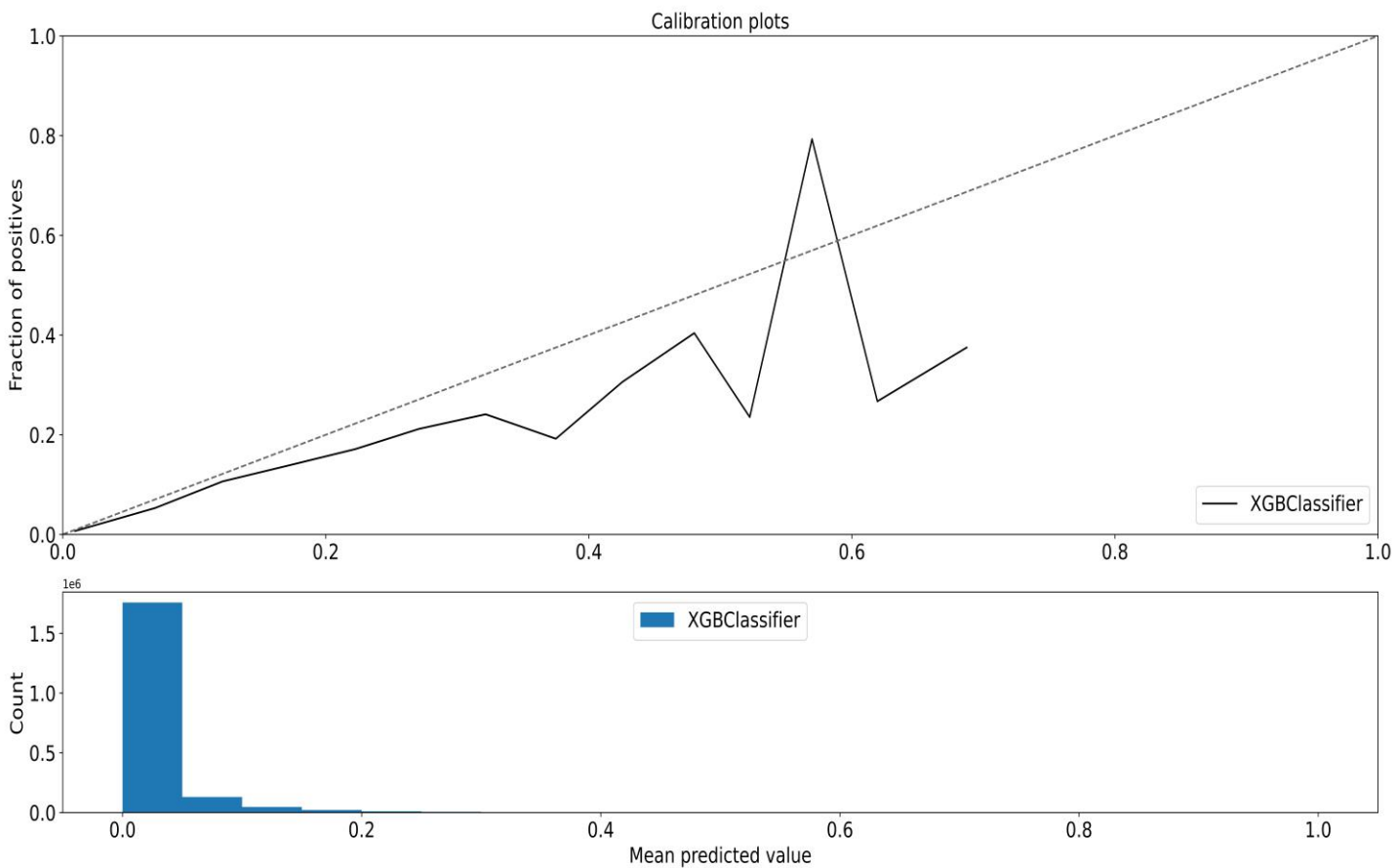
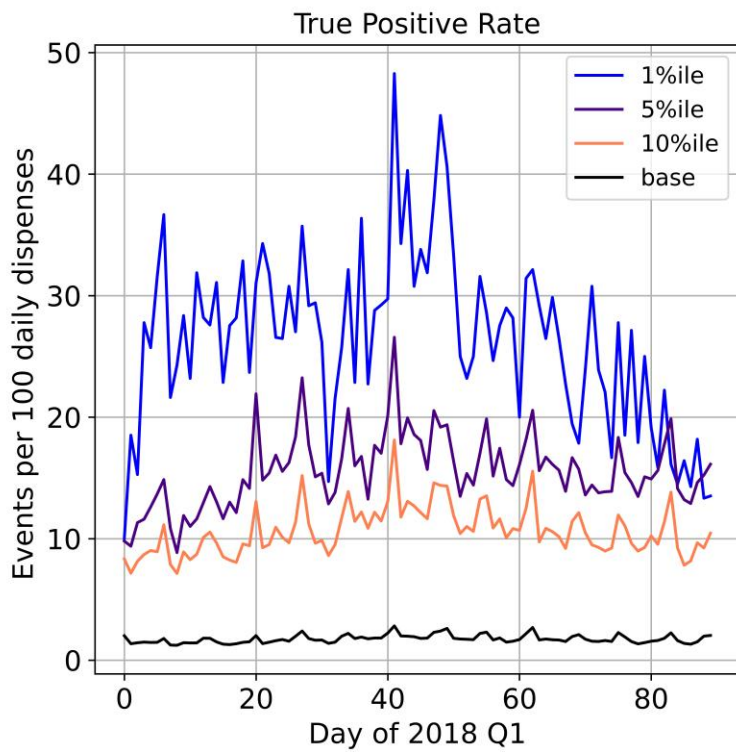
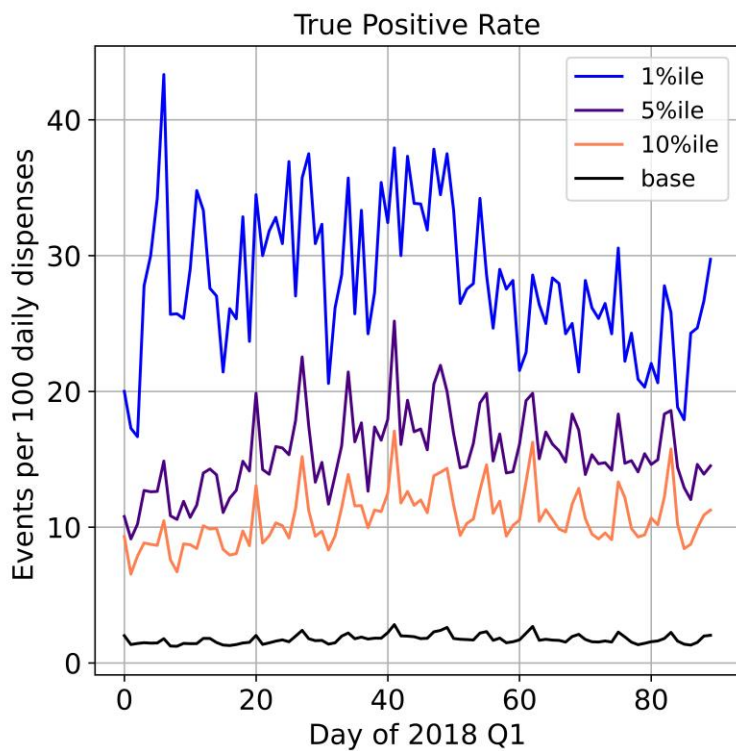


Figure 4. Simulation of a clinical workflow with daily uploads and events per 100 daily dispenses by risk percentiles using 2018 Quarter 1 (Q1) data for logistic regression (A) and XGBoost (B) classifiers.

(A) Logistic Regression (L1)



(B) XGBoost



view only

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 chi² 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

1
2
3 functions has shown to be very effective recently in many application domains such as natural
4 language processing, computer vision, genomics, computer games and health[2]. Neural
5 networks come in many flavors learning nonlinear mapping of different types of data such as
6 Convolutional NNs being most effective with images and Recurrent NNs for time series and
7 language data. Identifying the most effective neural network structure is one of the difficult and
8 the most time-consuming aspect of applying neural networks to new application domains and
9 data. Generally, neural networks try to exploit the relationships in the raw unstructured data (eg:
10 image and text) presented to the network but with more structured data such as health records
11 and ICD codes learning relationships is much complex. Our neural network models are mainly
12 based on densely connected hidden layers with ReLu[6] activation function. We used the cross-
13 entropy loss for the binary classification Adam optimizer. We used a simple feed forward
14 network using Sklearn MLP classifier with hyperparameter tuning for the NN.
15
16
17
18

19 **Boosted Learning Algorithms**

20
21 Boosting is a process to ensemble multiple base learning algorithms to arrive at better overall
22 performance than any individual base learner[1]. In contrast to independently building multiple
23 models from the subsets of the data, boosting re-weights the training data every time a model is
24 learned for future models. This weighting happens to give more preference to currently
25 misclassified data points in the next round compared to the correctly classified data points.
26 Therefore future learners try to do better on the misclassified data points leading to a collection
27 base learners having a better-combined prediction. This process is sequential so each base
28 learner is dependent on the output of the previously trained model (it is worthy to note XGBoost
29 provides a parallel tree boosting alternative). In our work, we have experimented with several
30 boosting meta-learning algorithms such as XGBoost[7], AdaBoost[5] and GBM[5]. XGBoost uses
31 a variant of trees as the base learner whereas AdaBoost (from Sci-kit learn) can use many ML
32 algorithms as base learners. GBM uses logistic regression by default as the base learner. We used
33 all 3 types of boosting with tuned hyperparameters for comparison.
34
35
36
37
38

39 **Naive Bayes**

40
41 Naive Bayes is based on the Bayes theorem with a strong independence assumption between the
42 covariates[1]. This assumption helps in building a simple probabilistic model for learning and
43 inference. Naive Bayes coefficients scale linearly with the number of covariates making this a
44 suitable model for high-dimensional data. We used Naive Bayes as a simple baseline learning
45 algorithm for comparison.
46
47
48

49 **Support Vector Machines (SVM)**

50
51 SVMs[4] are maximum margin classifiers optimizing for learning a hyperplane having the
52 maximum distance away from each of the class data points[1]. SVM is a linear classifier but with
53 the kernel trick to map the inputs to the higher dimensional space, it can learn nonlinear decision
54 boundaries in the input space. SVMs are very effective binary classifiers with the kernel trick[1].
55 With larger datasets, SVMs tend to become more computationally intensive.
56
57
58

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 χ^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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For peer review only

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 χ^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

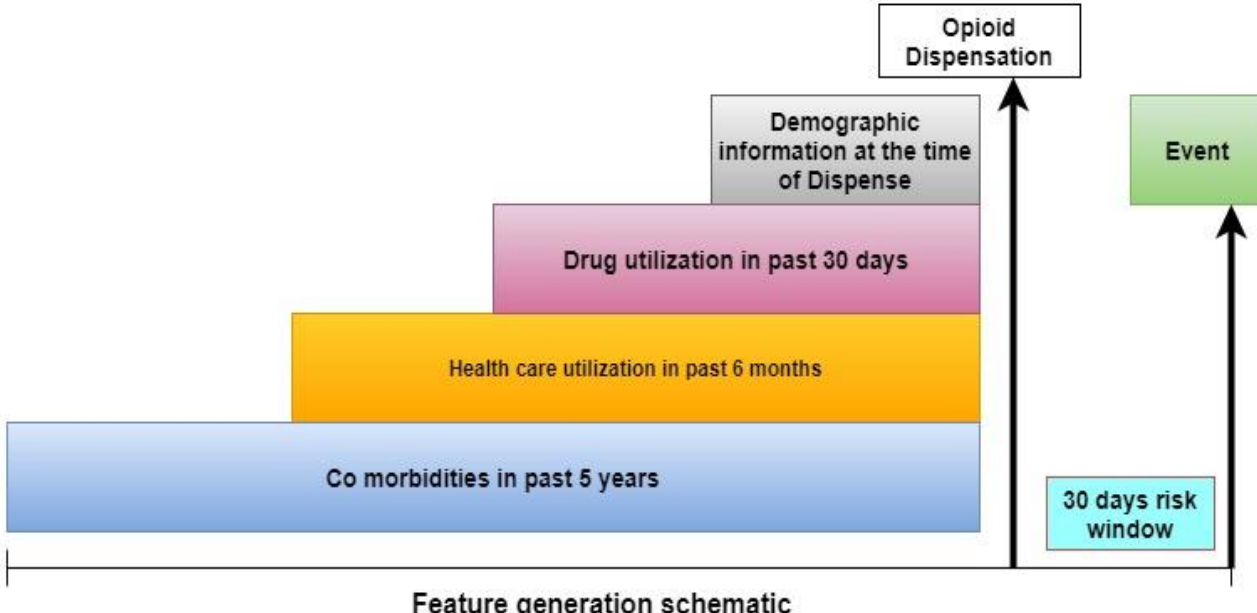
eTable 6. Discrimination performance using area under the receiver operating characteristic curve (AUROC) of various ML algorithms using all features (demographics, health utilization, prescription history, co-morbidities). Training and validation were done using 2017 data (n=393,979); another independent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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

eFigure 1. Schematic of study design and feature generation



view only

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

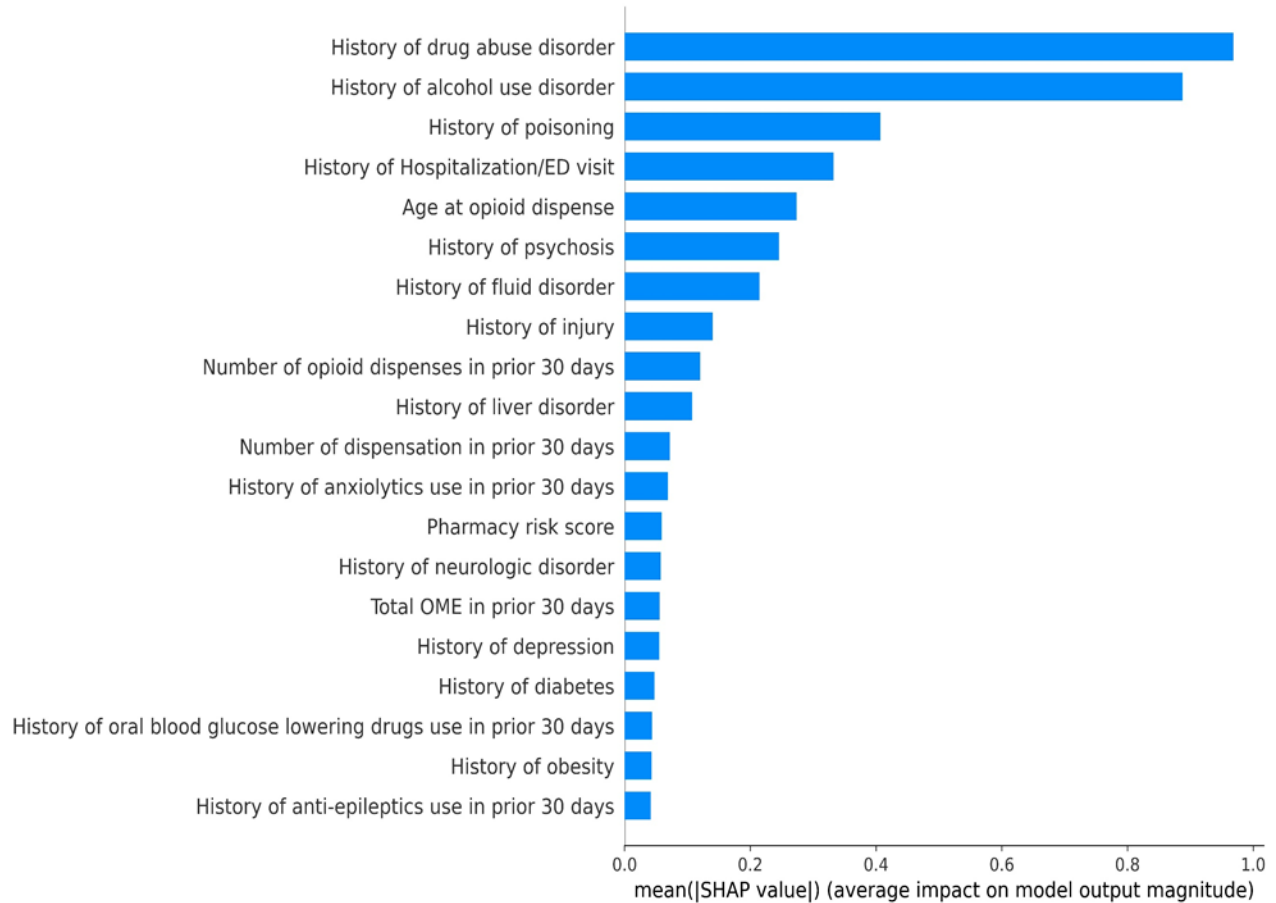
Note: Logistic regression used L1 (lasso) parameter regularization; BZD- benzodiazepine; OME- oral morphine equivalents; ED: emergency department

*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

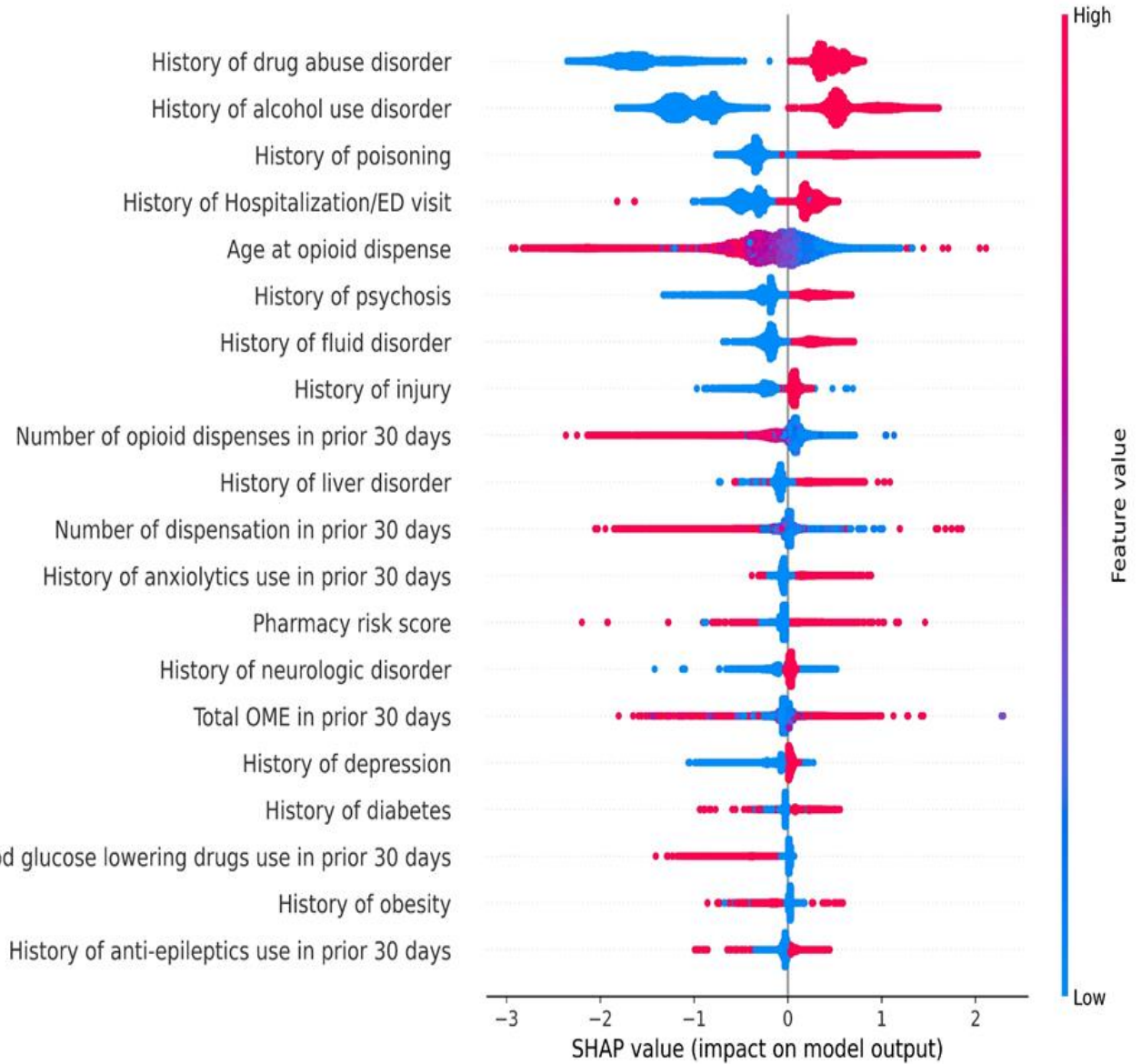
eFigure 3. SHAP values and feature impact of the XGBoost classifier using the 2018 validation set to describe “associations” between features and the outcome. Features with the most impact on the model with drug abuse ranked highest (A); tornado plot illustrating feature impact (B); explaining the prediction 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

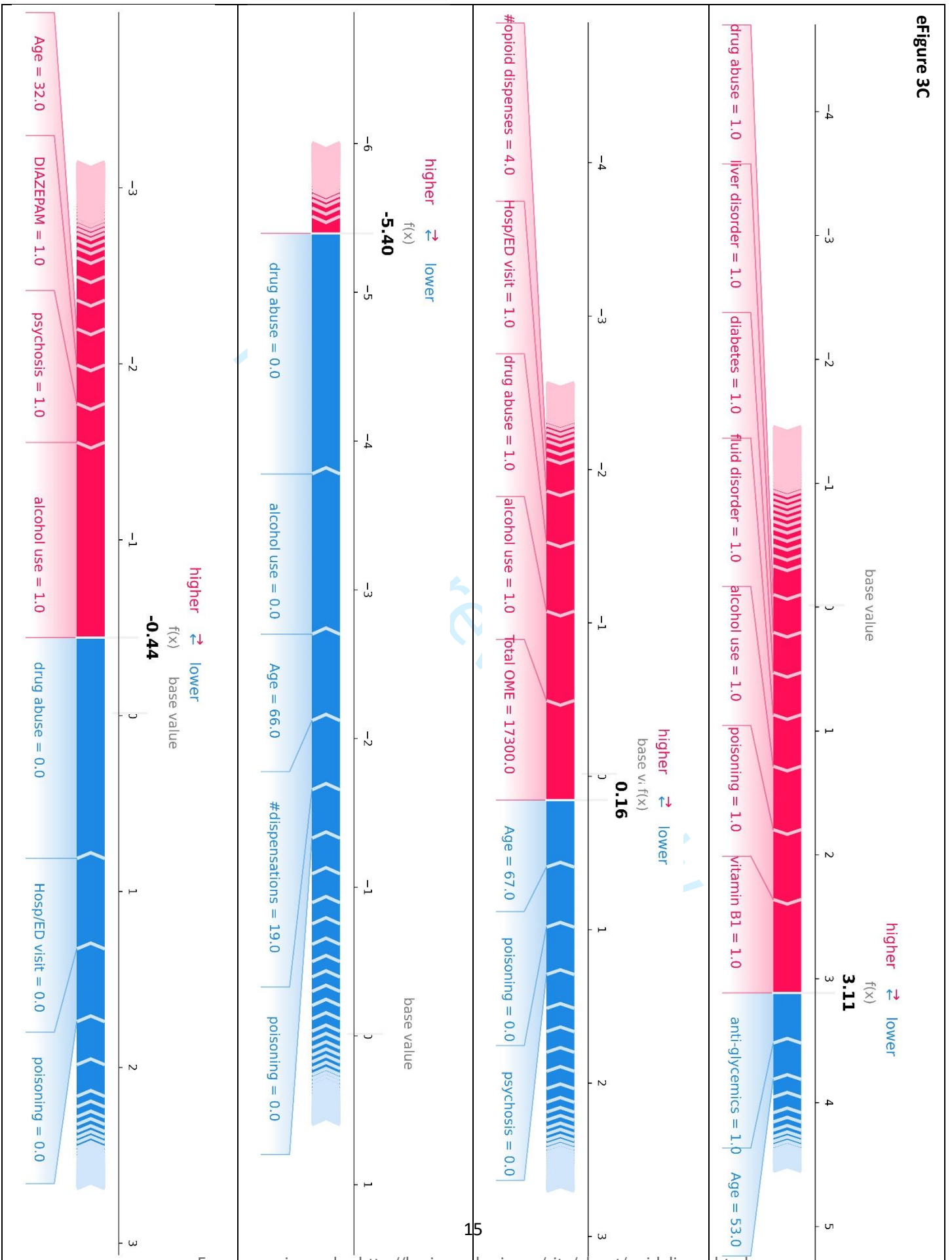
1
2
3
4 (C)
5

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

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

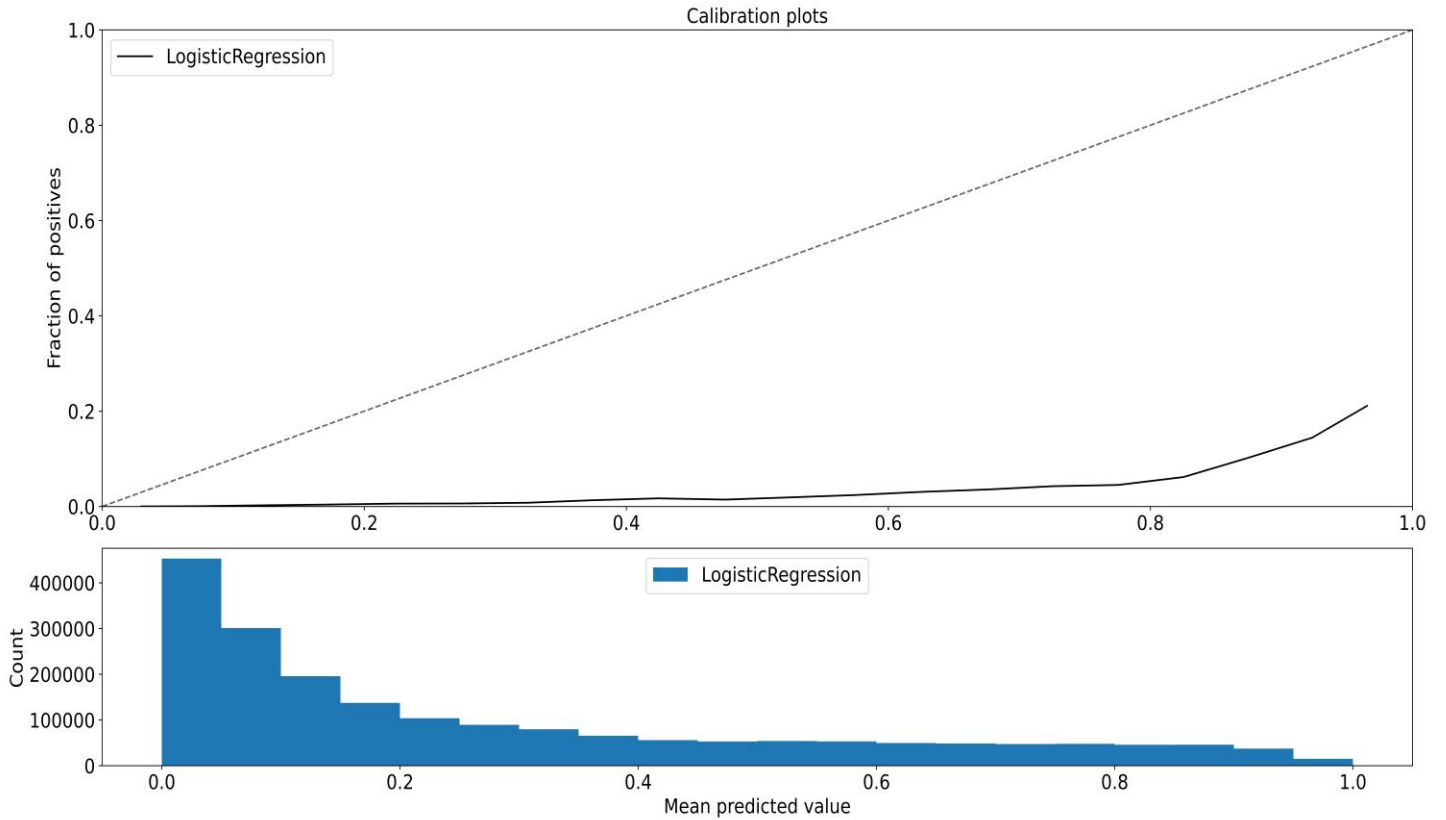
21 **Note-** drug abuse: drug abuse disorder; poisoning: history of poisoning; vitamin B1: vitamin B1 in prior 30 days;
22 anti-glycemics: anti-glycemic agents in prior 30 days; age: age at opioid dispensation; # opioid dispenses: number
23 of opioid dispensations in prior 30 days; Hosp/ED visit: history of prior hospitalizations and/or emergency visits in
24 past 6 months; Total OME: total oral morphine equivalents in prior 30 days; DIAZEPAM: history of diazepam use in
25 prior 30 days.
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

efigure 3C



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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).