Predicting Adverse Outcomes in At-Risk Populations with Machine Learning Methods

by

Vishal Sharma

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Epidemiology

School of Public Health

University of Alberta

Abstract

Canada has a publicly funded health system and heterogeneous population. There are segments of this population which account for substantial health care utilization and adverse outcomes. Machine learning (ML) approaches can assist public health intervention programs to mitigate health system costs and improve patient outcomes, particularly for segments within Canadian society that qualify as at-risk. The specific ones to be studied in this PhD are prescription opioid users, older adults taking benzodiazepines, and people with heart failure (HF). Identifying high risk individuals within these segments using ML methods trained on administrative health data as well as assessing prediction performance and value to inform health system planners are the objectives of this PhD program. This PhD studied outcomes related to admissions and deaths and presented findings on potential cost savings of ML assisted programs.

The main findings in this thesis were:

1. Machine-learning classifiers, especially incorporating hospitalization and physician claims data, have better predictive performance compared to guideline or prescription history only approaches when predicting 30-day risk of adverse outcomes pursuant to an opioid dispensation. Prescription monitoring programs and health departments with access to administrative data can use ML classifiers to effectively identify those at higher risk compared to current guideline-based approaches,
2. Despite predicting readmissions in patients with HF better than the LaCE, even the best ML model trained on administrative health data (XGBoost) did not provide substantially informative prediction performance as it only generated a moderate shift from pre to

post-test probability. Health systems wishing to deploy such a tool should consider training ML models with additional data. Adding other techniques like Natural Language Processing, along with ML, to use other clinical information (like chart notes) might improve prediction performance,

3. Developing ML models using only administrative health data may not provide health regulators with sufficient informative predictions to use as decision aids for potential interventions, especially if considering daily or quarterly classifications of benzodiazepine risks in older adults. ML models may be informative for this context if yearly classifications are preferred. Health regulators should have access to other types of data to improve ML prediction, and
4. Prescription drug monitoring programs can use ML classifiers to identify patients at risk of adverse outcomes from opioids and potentially reduce health-care costs by intervening on high-ranked predictions. Better access to available administrative and clinical data could improve the prediction performance of ML classifiers, especially if probability thresholds are important, and thus expand opioid stewardship efforts and further reduce costs.

In conclusion, the findings suggest that ML methods may demonstrate value in opioid stewardship programs with limited benefits in predicting adverse outcomes in older adults taking benzodiazepines and readmissions in people with HF. Health systems wishing to integrate ML into their program planning may benefit from additional sources of data to train ML models. Data governance, bias and ML transparency are key issues requiring future research.

Preface

This thesis is an original work by Vishal Sharma. The studies presented in this thesis received research ethics approval from the University of Alberta Research Ethics Board under the following approvals: Pro00083807_AME1, Pro00097809, Pro000838807_AME6 and Pro00083807.

A version of Chapter 3 has been published as Vishal Sharma, Vinaykumar Kulkarni, Dean T. Eurich, Luke Kumar, and Salim Samanani, "Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada" *BMJ Open,* 2021;**11:**e043964.

A version of Chapter 4 has been published as Vishal Sharma, Vinaykumar Kulkarni, Finlay A. McAlister, Dean T. Eurich, Shanil Keshwani, Scot H Simpson, Don Voaklander, Salim Samanani, "Predicting 30-Day Readmissions in Patients with Heart Failure Using Administrative Data: A Machine Learning Approach", *Journal of Cardiac Failure*, 2021 Dec 20. DOI: https://doi.org/10.1016/j.cardfail.2021.12.004

A version of Chapter 5 has been submitted for publication as Vishal Sharma, Tanya Joon, Vinaykumar Kulkarni, Salim Samanani, Scot H Simpson, Don Voaklander, Dean T. Eurich, "Predicting 30-day risk from benzodiazepine/Z-drug dispensations in older adults using administrative data: a prognostic machine learning approach".

A version of Chapter 6 has been published as Vishal Sharma, Vinaykumar Kulkarni, Ed Jess, Fizza Gilani, Dean T. Eurich, Scot H Simpson, Don Voaklander, Michael Semenchuk, Connor London, Salim Samanani, "Development and Validation of a Machine Learning Model to Estimate Risk of Adverse Outcomes Within 30 Days of Opioid Dispensation", JAMA Network Open, 2022 Dec 27 5(12). DOI: 10.1001/jamanetworkopen.2022.48559

Vishal Sharma was responsible for study design, analyses and manuscript writing for the above studies. All others contributed to study design, analyses and final review of manuscripts for the above studies.

Acknowledgments

The following acknowledgments must be made to recognize significant contributions to this PhD program:

- Dr. Dean Eurich: supervisor
- Drs. Scot Simpson, Don Voaklander, and Salim Samanani: co-supervisors
- OKAKI staff: machine learning expertise and notably, Vinaykumar Kulkarni
- MITACS Accelerate Funding

Contents

Abstract ii
Preface iv
Acknowledgments v
Contents vi
List of Tables ix
List of Figures xiii
Chapter 1: Introduction 1
Statement of the Problem 1
Study Objectives 2
Thesis Submitted for Partial Fulfillment of PhD 3
References for Chapter 1 5
Chapter 2: Machine Learning Background 8
Definition of ML 8
Patterns of AI 8
ML Process 9
ML Spectrum 12
Types of ML Algorithms 12
ML Methods 12
Data and ML 13
Predictors 15
Outcomes 15
Missing Data 16
Evaluation of ML Classifiers in Population Health Settings 17
ML Interpretability 19
ML and “Traditional” Epidemiological Approaches 19
Why use ML 25
Key Issues with ML 25
Summary 28
References for Chapter 2 29
Chapter 3: Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada 33
Introduction 35

Methods 36
Study Design and Participants 36
Measures and Outcome 37
Predictor Candidates for ML Models 38
Statistical Analyses and Machine-Learning Prediction Evaluation 39
Patient and Public Involvement 40
Results 40
Patient Characteristics and Predictors 40
Machine-Learning Prediction Performance 41
Calibration 41
Discussion 42
Appendix to Chapter 3 52
References for Chapter 3. 68
Chapter 4: Predicting 30-day readmissions in patients with heart failure using administrative data: a machine learning approach 71
Introduction 73
Methods 74
Study Design, Setting and Participants 74
Data Sources 74
Measures and Outcome 75
Machine Learning Methods 76
Predictors 77
Missing Data and Outliers 77
Analysis and Prediction Evaluation 78
Results 80
Discussion 82
Appendix to Chapter 4 92
References for Chapter 4 109
Chapter 5: Predicting 30-day risk from benzodiazepine/Z-drug dispensations in older adults using administrative data: a prognostic machine learning approach 115
Introduction 117
Methods 118
Study Design, Setting, and Participants 118
Data Sources 118

Measures and Outcomes 119
Predictors and ML Methods 120
Missing Data 121
Analysis and Prediction Evaluation 121
Patient and Public Involvement 124
Results 124
Discussion 126
Appendix to Chapter 5 136
References for Chapter 5 154
Chapter 6: 30-day risk from prescribed opioids: creating a risk predictor for prescription drug monitoring programs using a machine learning approach 159
Introduction 161
Methods 162
Study Design, Setting and Participants 162
Data Sources 162
Measures and Outcomes 163
Predictors and ML Methods 164
Analysis and Prediction Evaluation 165
Cost Analysis 167
Patient and Public Involvement 168
Data availability 168
Results 168
Cost analysis 171
Discussion 171
Appendix to Chapter 6 183
References for Chapter 6 198
Chapter 7: Discussion 202
Future Research 206
Conclusion 209
References for Discussion 210
Bibliography 211

List of Tables

Table 3.1. 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.

Table 3.2. 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.

Table 3.3. 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.

Table 3.4. Diagnostic codes used to exclude patients who had cancer, were pregnant, or were under palliative care.

Table 3.5. Diagnostic codes used to identify the defined study outcome from emergency visit, hospitalization and death data.

Table 3.6. Baseline characteristics of study patients (n=392,979). Co-morbidities were determined using Elixhauser criteria. All p-values in the chi^2 test of independence were <0.001 unless otherwise indicated.

Table 3.7. Characteristics of study participants between training and validation groups using 2017 data.

Table 3.8. Candidate predictors used to train ML algorithms.

Table 3.9. 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).

Table 4.1. Discrimination performance of ML models and LaCE score using AUROCs.

Table 4.2. Prediction metrics for the XGBoost classifier across predictive thresholds.

Table 4.3. Predictors used for model training (data dictionary).

Table 4.4. Admission statistics and main reason for admission.

Table 4.5. Characteristics of those who were readmitted and those who were not. This descriptive analysis was done at the patient level in which patients could experience multiple hospitalizations each.

Table 4.6. Characteristics of patients in development and validation sets. This descriptive analysis was done at the patient level in which patients could experience multiple hospitalizations each.

Table 4.7. Distribution of missing laboratory data.

Table 4.8. Prediction metrics for the LaCE score.

Table 4.9. AUROCs for various combinations of feature groupings for the XGBoost classifier.

Table 4.10. Sub-group analysis stratified by type of heart failure.

Table 4.11. Distribution of missing laboratory data in the HF specific reference hospitalization subset.

Table 5.1. All-cause outcome prediction metrics stratified by top percentile of risk for the XGBoost classifier measured at the end of 2019 on both validation sets.

Table 5.2. Simulation metrics for predictions classified daily using XGBoost.

Table 5.3. Simulation metrics for predictions classified quarterly using XGBoost.

Table 5.4. Anatomical Therapeutic Chemical classification of BZRA molecules used for this study and candidate predictor categories used to develop the ML model.

Table 5.5. ICD-10 codes used to define our composite outcome.

Table 5.6. Co-morbidity characteristics of those who did and did not experience an all-cause outcome using 2018-2019 data.

Table 5.7. Co-morbidity characteristics of participants in the development and validation sets.

Table 5.8. Distribution of missingness among commonly ordered lab test results.

Table 5.9. Area under the receiver operating characteristic curve (AUROC) for various ML algorithms using the entire 2019 validation set.

Table 5.10. C-statistics for the all-cause outcome XGBoost classifier using fewer datasets and for the composite outcome and its individual components.

Table 5.11. All-cause outcome prediction metrics stratified by top highest risk dispenses for the XGBoost classifier measured at the end of 2019 on both validation sets.

Table 5.12. All-cause outcome prediction metrics stratified by absolute thresholds for the XGBoost classifier measured at the end of 2019 on both validation sets.

Table 6.1. XGBoost prediction metrics measured at the end of 2019.

Table 6.2. Prediction metrics simulated using daily and weekly measurements stratified by top dispenses using 2019 data. Participants were progressively excluded for 1 year if previously flagged as high risk.

Table 6.3. Prediction metrics simulated using daily and weekly measurements stratified by absolute probability thresholds using 2019 data. Participants were progressively excluded for 1 year if previously flagged as high risk.

Table 6.4. Anatomical Therapeutic Chemical classification of opioid molecules used for this study and candidate predictors used to develop the XGBoost classifier.

Table 6.5. Diagnostic codes used to identify the defined study outcome from emergency visit, hospitalization and death data.

Table 6.6. Characteristics of study participants (n=853,324) in the development and validation sets.

Table 6.7. eTable 4. Characteristics of study participants (n=853,324) according to outcome status.

Table 6.8. XGBoost prediction performance metrics based on threshold of predicted risk measured at the end of 2019.

Table 6.9. Prediction metrics simulated using daily and weekly measurements stratified by percentiles of predicted risk using 2019 data. Participants were progressively excluded for 1 year if previously flagged as high risk.

Table 6.10. Prediction metrics simulated using daily and weekly measurements stratified by top percentiles of predicted risk using 2019 data. Participants were NOT progressively excluded if previously flagged as high risk.

List of Figures

Figure 2.1. The relationship between ML, AI and data science. ML: machine learning; AI: artificial intelligence; DL: deep learning. Copyright © 2019, Cognilytica.

Figure 2.2. The 7 patterns of AI/ML. Copyright © 2019, Cognilytica.

Figure 2.3. The ML process defined by the Alberta Machine Intelligence Institute.

Figure 2.4. Knowledge-to-action framework from CIHR.

Figure 3.1. 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 3.2. 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 3.3. Calibration curve plotting observed vs. quantiles (deciles) of estimated risk for the XGBoost classifier using the 2018 validation dataset. Most counts (dispensations) were predicted to be lower risk.

Figure 3.4. 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.

Figure3.5. Schematic of study design and feature generation

Figure3.6. Feature importance from logistic regression and tree-based (XGBoost) classifiers using the 2018 validation set.

Figure3.7. Shapley values and feature impact in the XGBoost classifier using the 2018 validation set to describe “associations” between features and the outcome.

Figure 3.8. Calibration curve plotting observed vs. quantiles of estimated risk for the logistic regression (L1) classifier using the 2018 validation dataset. Most counts (dispensations) were predicted to be lower risk.

Figure 4.1. Patient flow diagram of study participants.

Figure 4.2. Area under receiver operating characteristic (A) and precision-recall curve (B) for the XGBoost model.

Figure 4.3. Calibration plot (A) and NPV (negative predictive value) vs predicted risk (B) for the XGBoost classifier.

Figure 4.4. Feature importance for the XGBoost classifier.

Figure 4.5. Timelines of data capture for candidate predictor categories.

Figure 4.6. Frequency of patients with multiple admissions.

Figure 4.7. SHAP plot indicating influence of various predictors on risk of readmission*

Figure 5.1. Study participant flow diagram used for developing and validating machine learning models. Note: NACRS: National Ambulatory Care Reporting System; DAD: Discharge Abstract Database; VS: Vital Statistics; PIN: Pharmaceutical Information Network; Claims: Physician Claims; Labs: Provincial Laboratory database; BZRA: benzodiazepine receptor modulator

Figure 5.2. Decision curve/net benefit analysis for all-cause outcomes and the XGBoost classifier for the entire validation set (A), and out of sample validation set (B). Note: “Prediction Probability” in the legend refers to XGBoost classifier predictions. “Treat all” refers to if all dispenses were intervened on and “treat none” refers to if none of the dispenses were intervened on.

Figure 5.3. Simulation of events per 100 daily benzodiazepine receptor modulator dispenses using the XGBoost classifier stratified by percentile of predicted risk. Baseline risk corresponds to the pre-test probability.

Figure 5.4. Variable importance graph of the XGBoost classifier predicting all-cause outcomes.

Figure 5.5. Feature generation timeline

Figure 5.6. Precision-recall and area under the receiver operating characteristic curves for the XGBoost classifier for the entire validation set (A) and the out of sample validation set (B).

Figure 5.7. XGBoost calibration plots for all-cause outcomes using the entire validation set (A) and the out of sample validation set (B).

Figure 5.8. All-cause outcome calibration for XGBoost across sex for entire validation set (A) and out of sample validation set (B).

Figure 5.9. Negative predicted value vs predicted probability thresholds for the XGBoost classifier predicting all-cause outcomes.

Figure 5.10. SHAP plot indicating influence of various predictors on risk of all-cause outcomes*.

Figure 6.1. Study participant flow diagram.

Figure 6.2. Decision curve analysis. Across most of the range of threshold probabilities, the XGBoost classifier had a lower net benefit than if none of the opioid dispenses were intervened on. Thus, acting on predicted probability thresholds for interventions may not be informative nor appropriate.

Figure 6.3. Simulation of predicting the top 20 riskiest opioid dispenses measured daily by progressively excluding participants previously flagged as high risk. The yellow line, what we predicted, represents a workload that the CPSA would have to consider (A); XGBoost classifier predicting daily risks in a simulation for the College of Physicians and Surgeons of Alberta stratified by top percentile categories of risk. Base risk is around 2.6% and represents the pre-test probability (B).

Figure 6.4. Variable importance for the XGBoost classifier. Variable importance bears no statistical meaning in terms of association.

Figure 6.5. Cost of admissions (hospitalizations and emergency department visits) using predictions ranked by percentiles (percentile categories are the cut-off points) for data in 2019 Quarter 2. Costs are associated with only true positive predictions and represent the maximum possible savings the machine learning classifier will realize at the given percentile threshold of prediction based on daily classifications by a health regulator (A); cost savings and cost of interventions stratified by intervention success rates for predictions ranked in the top 1 percentile for 2019 Quarter 2 (B). All dollar amounts are in Canadian currency.

Figure 6.6. Schematic of study design and feature generation.

Figure 6.7. The number of opioid dispenses per day during 2019 (average of 8241 dispenses per day) and the number of dispenses which resulted in our defined outcome (average of 212 or around 2.6% of dispenses which led to an event).

Figure 6.8. Discrimination performance (A) and precision-recall curve(B) of our XGBoost classifier using 2019 validation data.

Figure 6.9. Calibration plot of the XGBoost classifier using 2019 data.

Figure 6.10. Negative predicted value vs predicted probability using the 2019 validation set.

Figure 6.11. SHAP values and feature impact from the XGBoost classifier using the 2019 validation set to describe variable importance in relation to the outcome. Features are arranged from highest to lowest impact on prediction.

Figure 6.12. Cost savings and cost of interventions stratified by intervention success rate for 2019 Quarter 2 reported for both the top 5 and 10 percentiles of predicted risk. All costs are in Canadian dollars.

Figure 7.1. Implementation of ML into health-care applications. From *Lancet Oncol* 2019; 20: e262-273.

Chapter 1: Introduction

Statement of the Problem

Canada has a publicly funded health system and heterogeneous population. There are segments of this population which account for substantial health care utilization and adverse outcomes. Indeed, evidence from various jurisdictions has shown that small proportions of at-risk groups in our population consume the majority of health system resources and it is important to mitigate this effect to sustain our health care system and improve patient outcomes[1].

There are many of these at-risk patient groups in our population. Older adults and those with chronic diseases are prominent examples of groups at risk of high health care utilization [2-5]. Furthermore, emergent health care issues like the opioid crisis have also led to increased use of health care resources[6]. Consequently, all these groups have also been identified as high risk for adverse outcomes such as emergency department (ED) visits, hospital admission, or death leading to substantial burdens on both the health system and individuals[6-9]. Thus, health systems are interested in reducing adverse outcomes in these high-risk groups to save finite resources.

Given the impact of these at-risk groups on health care resources and adverse outcomes, interventions that focus on them could significantly improve patient outcomes and reduce health system spending, a common conclusion drawn from the published literature[1,10]. Accordingly, the first step in this proactive approach requires a mechanism to identify these high-risk groups at an individual level so that subsequent interventions can be targeted appropriately.

Supervised machine learning (ML) prediction is one such mechanism which can identify individuals in these groups at high risk of adverse outcomes[11]. This approach uses computer algorithms to build predictive models at a population level that can make use of large amounts of available data within a well-defined framework[12-15]. Furthermore, the ML process allows for simple deployment for population health and surveillance purposes. ML approaches are

increasingly being applied in population health studies to predict health outcomes, in which these population-level models are used to identify high-risk groups, direct preventative interventions and inform health system policy-makers[16]. However, at the time of writing, there are no consensus guidelines on how to assess the effectiveness of ML prediction in this setting, making ML prediction reporting unclear and inconsistent[16-18] illustrating a major knowledge gap.

In summary, ML prediction approaches can assist health systems wishing to reduce costs and adverse outcomes identify high-risk groups at an individual level to target mitigating interventions.

Study Objectives

The primary objective of this PhD program is to use ML prediction methods to identify high risk people within certain populations. We will predict outcomes related to ED visits, admissions, and/or death. Health jurisdictions could use these predictions for their respective intervention programs. Because there are no consensus reporting guidelines for ML prediction, our secondary objective is to present methods to assess the performance, utility and value of ML classifiers for use in population level studies and intervention programs.

In this PhD project, the at-risk populations we will study are opioid users, people with heart failure and older adults, all of which are high users of the health care system. Our studies will demonstrate the capabilities of ML prediction classifiers in these at-risk populations. Specifically, we will develop and validate ML classifiers that attempt to predict 30-day risk of adverse outcomes pursuant to an opioid dispensation, 30-day risk of readmission after hospital discharge in people with heart failure and 30-day risk of adverse outcomes pursuant to a benzodiazepine dispensation in older adults. Finally, we will develop and validate a ML classifier based on our previous work which can assist the opioid stewardship mandate of the College of Physicians and Surgeons of Alberta to identify at-risk Albertans who are prescribed opioids.

To accomplish our secondary objective, we will assess the ML classifiers using metrics commonly used in clinical prediction literature which is not commonly done in ML prediction

studies[18-22]. Furthermore, we will conduct a simple cost analysis to simulate potential savings to a health system in our final primary objective study.

Thesis Submitted for Partial Fulfillment of PhD

This thesis consists of a review of ML methods and background (Chapter 2). It is followed by four studies (Chapters 3, 4, 5, and 6) designed to address the study objectives.

Chapter 2 consists of material retrieved from the literature that provides a general background on ML. This section provides details ranging from the definition and types of ML, different ML algorithms, benefits of ML prediction, how to assess ML prediction and limitations.

In Chapter 3, results of the first study are presented. A ML classifier was created and assessed to predict risks from prescribed opioids. The rationale behind this study is that Canada has among the highest rates of opioid prescribing in the world and part of the response to address the consequences of this is to endorse safe use guidelines and opioid stewardship via prescription drug monitoring programs[6,7,23,24]. This study was published in BMJ Open (doi:10.1136/bmjopen-2020-043964).

In Chapter 4, results of the second study are presented. Here, a ML classifier was developed and validated to predict risk of readmissions in the heart failure population in Alberta, Canada. Heart failure is a chronic disease identified as having high rates of potentially avoidable readmissions resulting in substantial burdens to both health systems and individuals[4,9,25]. The work from this study was published in the Journal of Cardiac Failure (doi: https://doi.org/10.1016/j.cardfail.2021.12.004).

In Chapter 5, results from the third study are presented. A ML classifier was constructed to predict risk of adverse outcomes in older adults pursuant to a benzodiazepine dispensation. The rationale for this work is that older adults are prescribed high amounts of benzodiazepines which carries substantial risks[26-28]. A version of this work has been submitted for publication and is under review.

In Chapter 6, results from the fourth study are presented. In this work, a ML classifier was developed specifically for and with the College of Physicians and Surgeons of Alberta

(CPSA). The rationale behind this study is that the CPSA wanted a proof-of-concept ML classifier which could assist with its opioid stewardship mandate. A simulation was performed to demonstrate potential cost savings to the health system based on ML predictions. A version of this work is currently being considered for publication.

In Chapter 7, the final chapter, general discussion and conclusions are presented. This includes an overview of the research, summary of results from the projects, discussion on the strengths, limitations and importance of the research, and directions for future research.

References for Chapter 1

1. Chechulin Y, Nazerian A, Rais S, Malikov K. Predicting patients with high risk of becoming high-cost healthcare users in Ontario (Canada). *Healthcare Policy.* 2014;9(3):68.
2. Curtis LJ, MacMinn WJ. Health Care Utilization in Canada: Twenty-five Years of Evidence. *Canadian Public Policy / Analyse de Politiques.* 2008;34(1):65-87.
3. Ploeg J, Matthew-Maich N, Fraser K, et al. Managing multiple chronic conditions in the community: a Canadian qualitative study of the experiences of older adults, family caregivers and healthcare providers. *BMC Geriatrics.* 2017;17(1):40.
4. Jencks SF, Williams MV, Coleman EA. Rehospitalizations among Patients in the Medicare Fee-for-Service Program. *New England Journal of Medicine.* 2009;360(14):1418-1428.
5. Alberta Health Services. *AHS Report on Performance FY 2017-18. Unplanned Medical Readmissions.* 2018.
6. O'Connor S, Grywacheski V, Louie K. At-a-glance - Hospitalizations and emergency department visits due to opioid poisoning in Canada. *Health Promotion and Chronic Disease Prevention in Canada.* 2018;38(6):244-247.
7. 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.
8. Orpana HM, Lang JJ, Baxi M, et al. Canadian trends in opioid-related mortality and disability from opioid use disorder from 1990 to 2014 through the lens of the Global Burden of Disease Study. *Health Promot Chronic Dis Prev Can.* 2018;38(6):234-243.
9. Canadian Institute for Health Information. All-Cause Readmission to Acute Care and Return to the Emergency Department. 2012.
10. Gawande A. The hot spotters. *The New Yorker.* 2011;86(45):40-51.
11. 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.
12. Bastanlar Y, Ozuysal M. Introduction to machine learning. *Methods in molecular biology (Clifton, NJ).* 2014;1107:105-128.
13. 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.

14. Alberta Machine Intelligence Institute. Machine Learning Process Lifecycle. In:2019.
15. Shah NH, Milstein A, Bagley P, Steven C. Making Machine Learning Models Clinically Useful. *JAMA.* 2019;322(14):1351-1352.
16. Morgenstern JD, Buajitti E, O'Neill M, et al. Predicting population health with machine learning: a scoping review. *BMJ Open.* 2020;10(10):e037860.
17. 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.
18. Luo W, Phung D, Tran T, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res.* 2016;18(12):e323.
19. 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.
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. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *Bmj.* 2016;352:i6.
22. 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.
23. 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.
24. 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.
25. Frankl SE, Breeling JL, Goldman L. Preventability of emergent hospital readmission. *The American journal of medicine.* 1991;90(6):667-674.
26. Cunningham CM, Hanley GE, Morgan S. Patterns in the use of benzodiazepines in British Columbia: examining the impact of increasing research and guideline cautions against long-term use. *Health Policy.* 2010;97(2):122-129.
27. Weir D. Benzodiazepine Receptor Agonist & Z-Drug Dispensations in Alberta: A Population - Based Descriptive Study2015.

28. ChooseWiselyCanada. The Canadian Geriatrics Society has developed a list of 5 things physicians and patients should question in geriatrics [Internet]. https://choosingwiselycanada.org/geriatrics/.

Chapter 2: Machine Learning Background

Definition of ML

The definition of ML varies by source and is vague, especially since the term ML is intermingled with data science and artificial intelligence (AI) (Figure 2.1). The term "AI" is used rather loosely and generally refers to the broad discipline of creating intelligent machines such as self-driving vehicles and digital personal assistants[29,30]. ML is a subset of AI and refers to systems (computer algorithms) that can learn by themselves from data and essentially is about discovering patterns and learning from those patterns to provide value beyond just analysis [30-32]. Data science is the field that understands how to extract value from raw data[32]; ML is driven by data science[33]. Deep learning is a subfield of ML focusing on neural network algorithms which attempt to simulate the behaviour of the human brain using large amounts of data.

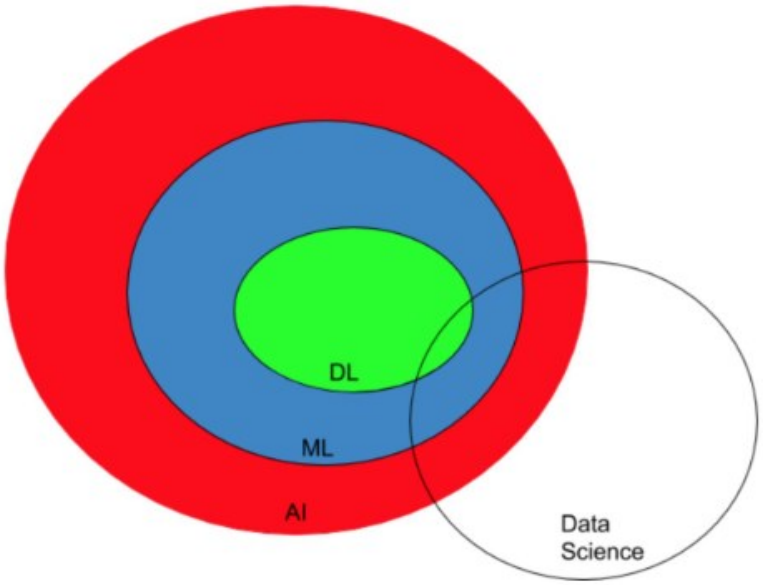


Figure 2.1. The relationship between ML, AI and data science. ML: machine learning; AI: artificial intelligence; DL: deep learning. Copyright © 2019, Cognilytica.

Patterns of AI

There are 7 patterns, or types of AI[32]. This PhD project is involved with only one pattern, namely, predictive analytics & decisions (see Figure 2.2). In this PhD program, ML classifiers will be constructed using administrative data and evaluated to eventually aid in

decision making in the public health arena. It is important to note that based on need, data and end-users, other patterns of ML could be followed in the public health world, such as natural language processing[34] (chart notes as a source of data) which falls under the pattern "conversation and human interaction".

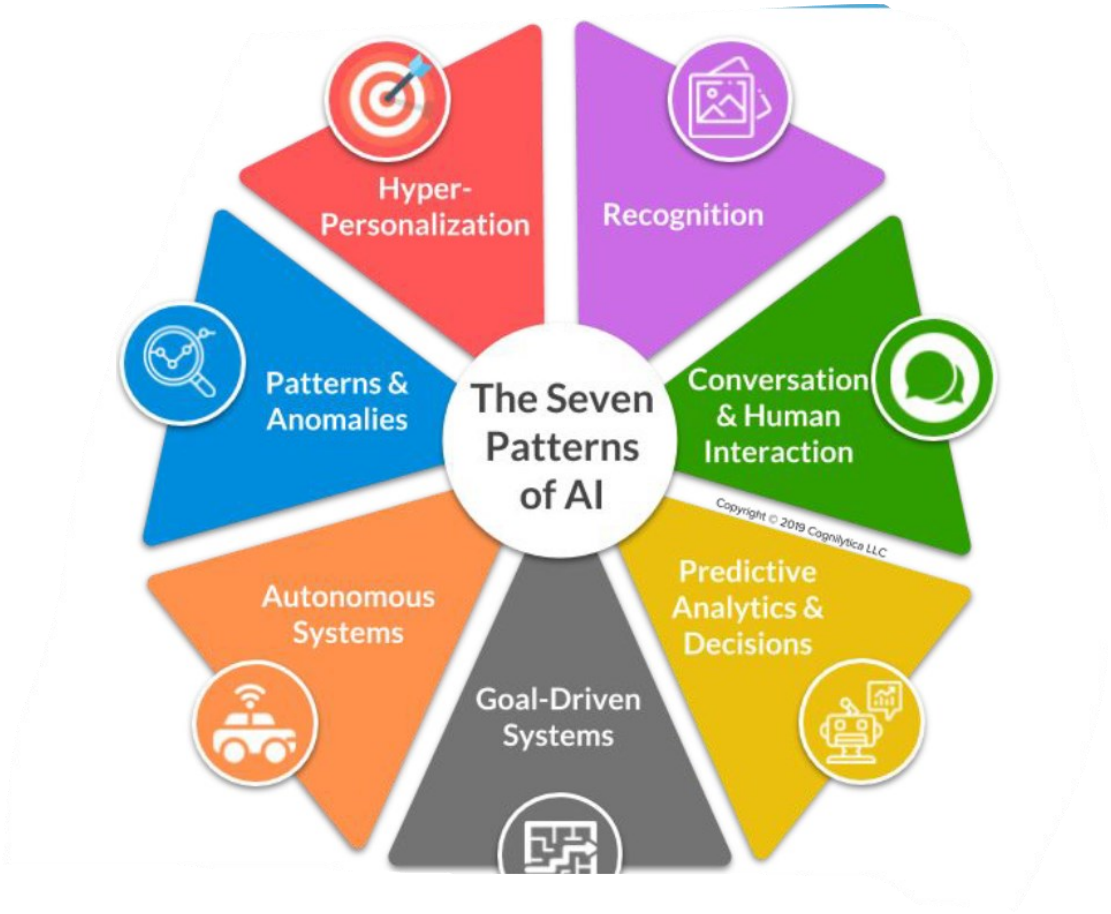


Figure 2.2. The 7 patterns of AI/ML. Copyright © 2019, Cognilytica.

ML Process

Defining ML also involves describing the ML process and ML in the context of precision medicine/public health. Various organizations have laid out a ML process lifecycle and share common elements. This PhD program followed the Alberta Machine Intelligence Institute's (AMII) model which focuses on business understanding, data acquisition & understanding, ML modeling & evaluation, and delivery (see Figure 2.3).

Figure 2.3. The ML process defined by the Alberta Machine Intelligence Institute.

This ML process lifecycle is very similar to the knowledge to action framework from the Canadian Institutes of Health Research (CIHR)[35] (see Figure 2.4). This is important because these PhD ML projects follow the ML process lifecycle which also make it aligned with CIHR's knowledge translation pathway. The common elements here are that ML classifiers are created based on public health need and input for the purpose of deployment into action.

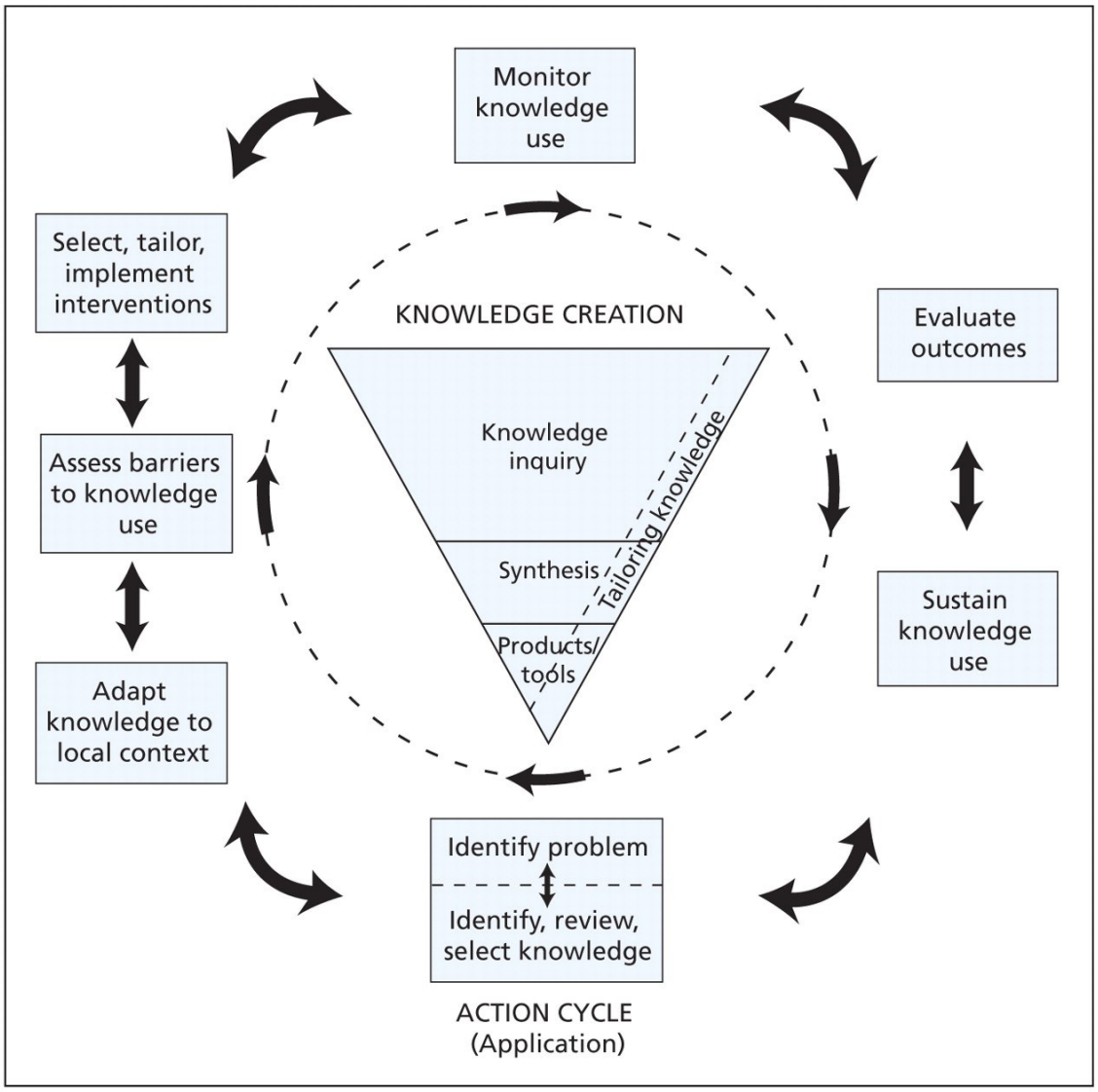


Figure 2.4. Knowledge-to-action framework from CIHR.

Finally, it is important to place ML in the context of precision medicine and public health. The general scheme for this was highlighted at an AI and population health conference[33]:

1. Identify a heterogeneous population
2. Collect individual person level data
3. Use ML to stratify some type of risk at the person level
4. Create a personalized treatment plan or a targeted public health intervention
5. Evaluate treatment or intervention effect

ML Spectrum

Another way, and more convenient, to define ML is with the concept of the ML spectrum. It is useful to describe ML as existing on a continuum between fully human guided to full data guided analysis and this trade-off, human specification vs data is defined as the ML spectrum[36]. This description of ML categorizes many existing clinical decision aids developed using "traditional statistical methods" and heavy human input (e.g., Framingham risk, CHADS score [congestive heart failure, hypertension, age>75, diabetes, stroke]) on the ML spectrum, albeit low. ML methods are a natural extension of traditional statistical approaches in which the decision aids are created more from data than by human input.

Types of ML Algorithms

A variety of ML algorithms for use in ML prediction have been studied in the literature[37] and general explanations of them are common[11,34]. A recent scoping review found that the most frequently used ML algorithms for prediction in population health were neural networks, support vector machines, linear models and tree-based methods[16]. The projects in this PhD will align with this finding by using these algorithms.

ML Methods

There are 3 types of ML schemes: unsupervised, supervised and reinforcement learning[11]. This PhD will only use supervised learning which involves training a model with input data (features) and its corresponding labels (outcomes). This ML scheme attempts to determine a relationship between the input data and label associated with the data.

Properly identifying the datasets used to construct ML classifiers is important because there is a lot of confusion regarding nomenclature. This PhD will be consistent with the medical ML community and use the following definitions[11]:

- Development set: a data set used for developing the ML model which is further split into the training and tuning sets,
- Training set: a subset of the development set that is used to develop the ML model where training is performed by updating the model parameters iteratively until the model optimally fits the data,

- Tuning set: a subset of the development set that is used to tune the hyperparameters of a model. In the general ML community, this set may be referred to as the validation set,
- Validation set: a data set that is independent from the training or tuning set. Validation sets are used to evaluate model performance before it can be deployed. The validation set should not be used for ML model training or hyperparameter tuning. In the general ML community, this set may be referred to as the test set.

Overfitting, in which a ML model fits the development dataset well but does not generalize to out-of-sample data, is a concern for all clinical prediction tasks[38]. In ML dialect, this is also known as the "bias-variance trade-off"; high bias implies incorrect predictions (underfitting) in the development and validation sets while high variance suggests the development dataset fits well to the ML model while the validation set does not[32]. ML methods include strategies to assess and minimize overfitting. K-fold cross validation is one such technique in which the development set is split into k groups where each of the k groups is used as a tuning set while the other groups are used for training. Hyperparameters, which are unique to each ML algorithm, are fixed parameters determined before an algorithm is trained and are optimized in each of the k test sets[11,32,39]. Regularization is used in conjunction with k-fold cross validation and involves early stopping (terminating the training process before overfitting occurs for neural networks), ensemble technique (combining and averaging multiple ML model outputs) and parameter regularization (shrinkage of parameters; e.g., lasso or ridge logistic regressions)[11,40]. All these techniques will be used in this PhD study; hyperparameters will be tuned using k-fold (k=10) cross validation. For each ML algorithm, the ML domain specialist will start the modeling process with the default set of hyperparameters provided. The optimal set of hyperparameters is where bias and variation are minimized such that the ML models are not overfitted to the development set data and are therefore assumed to be more generalizable to out-of-sample scenarios.

Data and ML

In the ML community, it is well known that the actual ML algorithms are quite simple and the complexity comes entirely from the data. Indeed, ML methods are a "data-first" approach. In fact, ML projects, including this PhD project, are essentially data management projects and it

is estimated that data preparation occupies at least 80% of the space in the ML process lifecyle[32]. The "quality" of the data is extremely important because even if the best predictive algorithm is suppled with inaccurate inputs, the result will be wrong predictions in the form of false positives and negatives[41].

The literature indicates that data falls into one of five categories and health systems should have access to all of them[34]. Broadly speaking, these data categories are measures of biology (e.g., genomic data sets), measures of context (e.g., geospatial data), administrative health data (e.g., electronic health records), personal monitoring (e.g., frequent device monitoring), and measures from effluent data (e.g., internet search term results). Data can be further delineated as structured (pre-defined data model) vs. unstructured (no pre-defined data model)[32]. Health systems could benefit by having access to this entire taxonomy of both structured and unstructured data[16,32,34].

A recent review found that most ML prediction studies in population health used structured, multi-linked administrative data[16] and this PhD program is no exception due to the available data in the province of Alberta. This PhD program will solely use administrative health data by linking data and outcomes from hospital records, emergency department visits, registry files, vital statistics, laboratory test results, pharmaceutical and physician claims histories.

The format, or organization, of the data used in ML model development can vary. Each line of data can be represented at the person or instance level, as is seen in the ML prediction literature[37,42]. In the former, each line is a unique person where in the latter, each line could be an admission or drug dispensation event. At the instance level, a person could be represented in multiple instances, creating correlations in the data. In some studies with multiple instances per person, the researchers arbitrarily picked the first occurrence to create a data set at the person level[42]. Still, others[43,44] have organized their ML data similarly to discrete-time survival analysis studies, a method convenient for longitudinal data collected in rolling windows at the person level. Nothing in the literature suggests that one type of data organization is superior to another in terms of prediction performance. Indeed, the data format must align with the research question being studied. This PhD program organized data at the instance level to align

with the research questions and did not arbitrarily drop instances to create a person level dataset.

Predictors

A review of the literature found that most population health ML studies focused on predictor features customary to clinical prediction models such as demographics and medical histories and that there was limited use of predictors from effluent data[16]. Although the reason may vary, in almost all cases it is simply because effluent data is not systematically captured by most health systems to be used in ML studies.

Predictor variables for the ML models in this PhD will include human derived ones informed by the literature and those directly obtained from the data. Several categories of candidate predictors to be used for ML training include demographics, co-morbidity history, healthcare utilization, and drug utilization.

Depending on predictor and data availability, data from 30 days to 5 years before the prediction window will be used to generate model features; 30 days to reflect the immediate nature of the risk and 5 years to fully capture co-morbidities and patient health care utilization.

Part of the ML process is to create a data dictionary which will evolve as the project progresses. This will detail all the predictors used in ML training. All predictors will come from physician claims, provincial registries, hospitalization and emergency department visits data sets.

Outcomes

As commonly seen in the literature[16], this PhD thesis will use measures of health care utilization as the outcome of interest for ML prediction. Our outcomes will be labeled to each instance in our dataset and will represent measures of public health and health system importance including admissions and death regarded as potentially preventable. Outcome data will come from hospital and ED admissions datasets.

Missing Data

Missing data is typically classified and handled in the context of descriptive and causal (association) studies where the desired result is an unbiased estimator of a population parameter with an accompanying standard error[45,46]. In these types of studies, missing data are classified as MCAR (missing completely at random), MAR (missing at random), and MNAR (missing not at random)[45,47]. How missing data is handled depends on assumptions about the missingness based on this classification. In order to handle, or impute, missing data, MCAR and MAR are assumed[48]. In most situations assuming MCAR and MAR, simple techniques for handling missing data (complete case analysis, overall mean imputation, missing indicator method) produce biased results[45], thus, MI (multiple imputation) is the method of choice for handling missing data[45,49]; however it also assumes data is MCAR/MAR, which rarely is true. Furthermore, the purpose of MI is to obtain more accurate standard errors of the estimate of interest[50]. Imputation methods with MNAR are not recommended because they are not well studied and the performance is unknown[47,51]. Mostly, missing data is neither MCAR nor MNAR, thus making missing data mostly MAR[45]. MAR data is difficult to comprehend because missing data can be considered random if its missingness is conditioned on the other covariates[47]. In the published literature, missing lab data is a mixture of MAR and MNAR[50,51]. In the context of descriptive and causal studies, imputation methods may yield varied results with this mixture[52].

Currently, missing data research has mostly been conducted in the context of parameter estimation, which is not directly relevant to prediction modeling studies[46]. For ML prediction, a different approach to handling missing data may be warranted to optimize predictive performance; the optimal way to handle missing data in prediction studies may be different than that of causal association or descriptive studies[46]. The MCAR, MAR, MNAR classification also applies only to parameter estimation studies, and not prediction studies. With ML prediction projects, the focus is on prediction and not association. This distinction is important because it changes the way missing data is considered and handled.

Most data, especially for ML projects, comes from administrative data bases and electronic health records, which will have missing data due to a mixture of MAR and MNAR. This would pose a challenge for parameter estimating studies, but may provide an opportunity

for prediction studies because the missingness may itself be a predictor of the outcome[46]. Researchers should consider if the missingness pattern is informative (may represent a latent variable). There are studies that show using missing indicators with or without MI improved prediction performance even when using MNAR data[46,53]. This contrasts with parameter estimating studies.

An important and overlooked issue with missing data and prediction studies is the distinction in handling missing data during model training, validating, and deployment (real world predicting)[46]. Even if multiple imputation and/or missing indicators resulted in powerful prediction metrics in the training and validation sets, the logistics of multiple imputation/missing indicators during real world prediction would not allow for deployment because in the real-world scenario, patient data would have to be linked to the training set, imputed, then re-modeled. Bottom line, the missing data approach used in ML prediction model development will not match the approach used when a model is deployed. One alternative to this problem of multiple imputation is to use the missing pattern method[54] where each pattern of missingness is modeled and at prediction time (deployment) imputation is not required. This method has its own issues related to data quantity, however, the researchers that developed it have offered solutions. Also, tree-based algorithms have an inherent capacity to handle missing data during model deployment[55]. At this point, this PhD project will not incorporate missingness into ML model development.

A scan of the literature found that very few researchers described how they handled missing data in their prediction modeling. In one study predicting surgical site infections[51], researchers only used logistic regression as their classifier (no tree-based modeling) and used imputation and missing indicator methods to handle missing data. They did not address the downstream issues of these methods during a potential deployment.

Evaluation of ML Classifiers in Population Health Settings

There are no established guidelines for evaluating and reporting ML prediction in public health settings. This makes assessing the utility and value of ML classifiers difficult for both readers and end-users, including health system planners. The published literature has pointed

out this knowledge gap. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement is a guideline for reporting studies of developing, validating, or updating a prediction model. A scan of the literature found that most ML studies did not use TRIPOD reporting guidelines in their work, lacked adequate detail on participants and disease distribution, and did not articulate if their validation groups correspond to the deployment setting[17]. Regarding prediction metrics, most studies reported discrimination performance (i.e., area under the reporting operating characteristic curve; c-statistic) and few reported calibration metrics[16,17]. Indeed, applications of ML in population health would benefit from increased assessments of calibration metrics[16].

Given this inconsistency in reporting on ML studies, researchers are starting to fill this void. A need for reporting guidelines is now acknowledged and progress is being made. A recent AI conference[33] attempted to characterize the issue and address challenges to public health translation. Evaluation challenges included issues related to ML generalizability (overfitting), interpretation and biased predictions. Challenges with ML translation into deployment relate to not modeling with the public health (clinical) setting in mind, lack of data, lack of guidelines for evaluation and ML interpretability. More emphasis on ML model validation in specific populations is required. Preliminary guidelines have been established while the long-awaited TRIPOD AI guidelines will be released soon[18,56].

Taking cues from the literature and identified knowledge gaps, this PhD's second objective attempted to create a template of ML prediction analysis which health system planners and other potential end-users would find informative. These metrics, briefly described in this section and further detailed in the subsequent study chapters, will help public health intervention programs decide if or when to intervene on individuals in a targeted population. These metrics include discrimination and calibration performance with a focus on comparing pre-test to post-test probabilities in various risk groups defined by rank or absolute threshold. For example, an intervention program could act on a high-risk group (e.g., a top 10 list) or on a certain threshold of predicted risk. Thresholds should ideally be ascertained by health system planners based on methods like decision curve/ net benefit analyis[16,21]. Through methods validated by the Canadian Institute for Health Information (CIHI)[57,58], this PhD thesis will assess

the value of ML classifiers by estimating potential savings of a ML assisted intervention program in a simulated deployment.

ML Interpretability

ML interpretability is an important issue when it comes to ML translation and so called "black boxes" are deemed unacceptable in health contexts[33,59]. Health system decision supports require some measure of transparency so that users can understand the basis for interventions aided by ML prediction[59]. With ML risk prediction, there can be a trade-off between model accuracy and intelligibility[60]. To optimize prediction accuracy, ML methods do not attempt to produce interpretable models thus, allowing them to handle large number of predictors common in most projects[18]. Although interpretability is not the primary focus of ML prediction, some ML algorithms (e.g., tree-based) are inherently more interpretable than others (e.g., neural nets)[34]. Based on these points, this PhD program will focus more on tree-based algorithms as much as possible and employ techniques to demonstrate ML interpretability involving variable importance, a rank ordering of variables that are most important to a ML model's prediction performance that has no statistical or causal meaning[50]. This piece of our analysis will provide health system planners or others with insight on how a ML classifier was influenced in its predictions for a public health intervention program.

ML and "Traditional" Epidemiological Approaches

Comparisons are frequently made between ML and so-called traditional epidemiological approaches. However, this is misleading because ML is an extension of traditional statistical approaches with a few key differences. It is important to make comparisons between the two on equal grounds, i.e., traditional approaches to individual risk prediction vs. ML risk prediction.

Accurate risk prediction is important for health systems, health providers and individuals for making optimal treatment decisions[19]. It allows for a shared decision-making process, reduces interventions in low-risk situations, and avoids intervention delays in high risk situations[19]. Risk is assessed using several methods[19]. Intuition is sometimes used but is not reliable in many cases. Population averages from observational studies are also used but can provide inaccurate estimates because of heterogeneity within and between populations.

Another limited method is the use of measures of association (estimation of relative risks or odds ratios) from risk factor studies that do not account for baseline risk. In this scenario, the reported relative effects associated with the risk factors under study can be misleading[19]. A more informative approach is incorporating several risk factors into a model in order to estimate absolute risk of an outcome in the individual patient. All but clinical intuition falls under the category of traditional epidemiology. The last method, estimating absolute risk of an event at the individual level, can be directly compared with machine learning approaches because both seek to accomplish the same objective, individual risk prediction.

Machine learning (ML) falls under the realm of artificial intelligence. Briefly, the ML approach uses a supervised learning scheme in which a data set containing historical information on people is labelled with the outcome of interest and is used to train a ML algorithm into a model[11]. After tuning and validating[11], the model (a mathematical function) is used to predict the outcome of interest at the individual level, usually by providing a probability score for an individual. Thus, ML prediction deals with probabilistic outcomes which is very useful to health systems for predicting clinical events such as hospitalizations or deaths. In this context, the focus of ML is individual risk prediction

Traditional epidemiological methods involve collection of data on a sample population and estimating population parameters to fit data to a model with probabilistic outcomes, a process very similar to ML. Instead of extrapolating from data obtained from a small number of samples to make estimates on a population, ML approaches use data at the population level to provide a real-world picture; this is a fundamental change from classical approaches which focuses on reducing effects of bias due to study design[29]. As mentioned earlier, the emphasis is individual risk prediction. Traditional epidemiological methods can estimate individual risk prediction by developing clinical decision rules (CDRs). A CDR is a clinical tool that estimates a patient's risk of an outcome by quantifying the individual contributions that various components of a patient's history make towards the prognosis of that patient[61]. $CHADS_2$[62] and Framingham scores are commonly used CDRs for estimating the risk of cardiovascular outcomes in individual patients. CDRs can be applied at the population level for public health programs or for surveillance.

There is considerable overlap of attributes between CDRs developed using traditional epidemiological methods and ML risk prediction since CDRs are considered an early incarnation of ML[11]. This overlap has led to the concept of the ML spectrum[36], where the traditional CDRs fall on the lower end (more human than machine specification of important predictors) of the spectrum and where the more complex ML algorithms (neural networks with no human specifications) fall into the higher end of the spectrum. Nevertheless, there are substantial differences between traditional epidemiologic and ML methods. The framework for comparing ML and traditional epidemiologic methods follows: model development including data and analysis issues, process definition, performance metrics, reporting, intelligibility/interpretability, and implementation/uptake.

CDRs and ML risk predictors must both follow three steps of development[19,61]. The first step is derivation of the CDR or ML risk prediction by identifying factors with predictive power using data and analyses. Data quality is at the heart of this step; data must be in a structured format before it can used to train CDR or ML models. Participant recruitment methods heavily influence both traditional epidemiology and ML methods and share the pitfalls of bias, which can then be perpetuated in their risk predictions. A commonly known fact is that all types of observational studies and traditional statistical modeling may be biased. ML algorithms may also be subjected to biases[63]. The biases include those related to missing data and patients, participant selection (selection bias) and underestimation, and misclassification and measurement error[63], all similar to the biases found in traditional epidemiology. Much published literature describes data bias in ML projects and some of these suggest that bias may be more pronounced and influential in the ML world[64]. One reason for this is that bias in ML projects may be harder to recognize due to "automation complacency" where users inadvertently ignore relevant information from non-ML sources[64]. The result of bias is the same for both traditional epidemiology and ML methods, an invalid estimate of risk. On a side note, the generalizability of both traditional epidemiology (observational studies) and ML risk predictions across different sub-groups is dependent on the representativeness of the included populations, missing data and outliers[64].

The other part of the derivation step is data analysis. Traditional CDRs are typically developed using linear models (e.g. logistic regression) to determine which predictors are influential in risk prediction and which can be omitted from the model[61]. Linear models are classified as a type of ML algorithm and as mentioned earlier, place CDRs on the low end of the ML spectrum because humans typically specify many of the model parameters. On the other hand, ML risk prediction methods include many other algorithms in addition to linear models. Some of these include deep learning algorithms (neural networks), random forests, gradient boosting machines and many others that are described in the literature[37]. Depending on how much human input is involved, these algorithms are placed much higher on the ML spectrum where some neural networks have no human specifications and are classified as fully intelligent[36]. One of the attractive attributes of ML risk prediction is that it can incorporate all available predictor variables in the algorithms to produce an estimate of risk however, this would reduce intelligibility as described later. The same could be done for traditional CDRs but would make them too cumbersome to use. Typically, ML methods use more features (predictor variables) than traditional CDRs in their risk prediction models[11]. Indeed, the "curse of dimensionality" associated with large data sets is somewhat eased by using ML methods[34]. The current hype about ML prediction is that the newer, more advanced algorithms (neural networks, boosting machines) are associated with better prediction estimates than the older linear models used by CDRs[11]. When the risk prediction literature is scanned, the results are mixed, with some studies showing the simpler linear models performing better than more complex ones and vice versa[65].

The second step in risk prediction development is validation, evidence of reproducible accuracy[19,61]. There are differences in the validation procedure between traditional epidemiology (CDRs) and ML methods. In traditional CDRs, the parameters (weights or coefficients from linear models) are generally derived from a single development/training set and then evaluated on one or more independent validation sets[11] using performance metrics. In ML risk prediction, although parameters are derived in a similar way from the data, there are additional hyperparameters, such as learning rate, that affect parameter estimation. These hyperparameters also need to be "tuned" using an additional "tuning" set that is usually an

extension of the development set and independent of the validation set[11]. Hyperparameters need to be tuned to reduce "overfitting" because they have a large effect on final risk estimates. Overfitting is a scenario in which the trained ML model is too specific towards the training data set and does not generalize well to other data sets[11]. Tuning hyperparameters falls under the domain expertise of ML engineers, as described later in the process section. Validation metrics in ML are like that of CDRs and include c-statistics, calibration to name a few.

The third step in the development of risk prediction is impact analysis[61]- does the risk estimator change behaviour, improve patient outcomes and reduce costs? Both traditional CDRs and ML risk prediction must execute this final step to be considered beneficial, or else no matter the accuracy, the risk prediction tool, either CDR or ML, will not be systematically used[61]. There are many validated CDRs and ML prediction models but only few are used in clinical settings[66]. Reasons for limited use may be related to provider unawareness, lack of interpretability of the model, and lack of impact studies[19]. The best way to asses the impact of CDRs and ML risk prediction tools is to randomize patients or institutions to the risk predictor or not and to follow up and measure the relevant variables and outcomes or to use controlled before and after studies[61].

The processes that define CDRs and ML risk prediction development are also different. The ML process has been well defined into a project management framework and many organizations, like AMII (Alberta Machine Intelligence Institute), have described it. The basic steps are: business understanding & problem discovery, data acquisition & understanding, ML modeling & evaluation, and delivery & acceptance[14]. Traditional CDRs do not have a formalized process of development like ML projects. Furthermore, underlying the ML process is an added area of domain expertise that requires the expansion of the research team to include ML programmers. Such domain expertise is not part of the traditional CDR research team.

Performance metrics and their subsequent reporting in peer reviewed arenas are almost identical for both traditional CDRs and ML risk prediction models. Published literature on ML risk prediction commonly report discrimination metrics such as sensitivity, specificity, predictive values, likelihood ratios, number needed to evaluate, and area under the receiver operator

characteristic curve (AUROC)[19,37,66]. CDRs report the same performance metrics like in the opioid risk tool[67]. Occasionally, ML risk prediction studies will estimate F1 scores, accuracy, and precision-recall curves[37]. Calibration analysis is also important to perform on both methods where observed to expected ratios are illustrated[19]. Guidelines for reporting on studies dealing with model derivation and validation are the same as well; the TRIPOD guideline statement is commonly used[22,68]. ML specific reporting guidelines are currently in the works.

One of the key differences between traditional CDRs and ML risk prediction deals with model intelligibility, or model interpretability. With ML risk prediction, there can be a trade off between model accuracy and intelligibility[60]. Supposedly more accurate models using neural nets are not intelligible to users and have a "black box" reputation while general linear models used in CDRs are more interpretable; ML risk prediction using linear models would also be considered more interpretable. Furthermore, the more complex ML algorithms may not necessarily produce more valid risk estimates [60,65]. The issue of interpretability is crucial because this is often the deciding factor when implementing a risk predictor, either from CDRs or ML[19]; deploying a ML risk predictor that uses a neural net model may be considered "too risky" because it is not easily understood by the users[60]. Also, many CDRs developed using logistic regression that are used in practise are easily interpretable with "lower accuracy". An example is the CHADS scoring system that is highly interpretable but with an AUROC 0.66-0.75[19]; many ML risk prediction tools have higher discrimination performance but are not used because of lack of interpretability[60]. Also, traditional CDRs have fewer risk factors to consider, making them easy to use while ML risk prediction models may have hundreds of features.

Finally, the implementation of traditional CDRs and ML risk predictors is also different. CDRs that are validated are easily implemented and used. Users can apply a CDR by consulting a risk table, calculator, or even mentally counting risk factors[11]. Implementing a ML prediction tool into the workflow is more complicated[11]. Computer programs to manage data pipelines have to be established, privacy rules have to be accommodated, and IT maintenance all have to be considered[11]. However, one benefit of ML risk prediction is that ML models can be updated frequently, often in real-time, to fit the local population. Traditional CDRs are static and risk calculators like the Framingham have issues with prediction performance and generalizability

when applied to different populations[69]. In these scenarios, CDRs like the Framingham cannot be quickly re-calibrated to the new population. ML risk prediction models are different since their accuracy can be improved over time because, as a response to changes in practice or patient population, ongoing data collection can lead to improved ML model performance[11].

Overall, there is no clear distinction between traditional epidemiological and ML methods because of the considerable overlap in development and function. CDRs are generally more interpretable for the time being, however, this may change as awareness of the capabilities of ML risk prediction increases. ML models are easily updated to adapt to new situations. In other words, traditional epidemiological methods may evolve to incorporate ML methods for the purpose of risk prediction.

Why use ML

So why use ML? In summary, ML risk prediction does offer some benefits over traditional risk predictors. ML methods allow for incorporation of large amounts of data and modeling of more complex and non-linear data; traditional techniques of regression require more human input to structure the data with underlying assumptions while ML algorithms derive structure directly from the data making fewer distributional assumptions and require less human input[16]. Advances in computing allow ML classifiers to include a larger number of predictors and be scaled up to the population level; they are not restricted to "pocket-card" sized risk calculators. Risk classifiers high on the ML spectrum require less human input and depend more on the data. These are all important factors when considering our strained health system. Furthermore, ML ideally allows the opportunity to continually monitor and learn from new data thus improving prediction performance over time.

Key Issues with ML

As described in the literature and presented in conferences, there are some key issues surrounding the use of ML by health systems. These issues are related to data, ethics, ML interpretability and reporting on ML performance.

Most ML prediction studies focus on features typical of clinical prediction models such as demographics and biomedical factors from administrative health records with limited, or no

use, of other types of data, either structured or unstructured; a reliance on this narrow approach is unlikely to fully leverage the benefits of ML prediction for health systems because key predictors may be absent[16,41]. A substantial amount of data of interest are not being held in administrative health databases but by industry and people[34,70]. Modern techniques allow for linkage of this siloed data and to break the cycle of relying solely on administrative health data; all categories in the taxonomy of data should be available for ML modeling[16,29]. Barriers to data sharing also add another layer of complexity[71]. The issue of data governance, along with the growing amount of data from wearable personal devices, are now recognized as major contributing factors to predict health system outcomes[71].

Further, data currently in use was never collected specifically for ML prediction. The large amounts of data "dumped" into ML development represent a mixture of local, regional, provincial and national level data[33,70]. However, this situation is improving as data collection is becoming more organized. Data sets not collected for ML use pose several issues: missing data, problems with anonymization, poor quality and bias (social factors)[71].

Bias in data is a major consideration for health systems interested in ML assisted programs as health data does not always include data from all patient populations, especially under-served ones[71,72]. The resulting predictions may lead to certain segments of a population to be excluded from the benefits of ML prediction. The ML modeling process itself can also be a source of bias leading to discriminatory prediction performance[63]. Adding to this is the complexity of ML methods which can hide the source of bias (e.g., bias in data, ML algorithm, participant selection) making it difficult to address[71]. Indeed, the literature is starting to describe scenarios of bias and discrimination related to these issues in ML assisted health system planning[72].

Issues related to ethics, privacy and consent also must be highlighted. Obtaining consent is complicated because large data sets are often used many times for different purposes by multiple users but this in itself is no excuse not to pursue consent where feasible[71]. Furthermore, data must be held secure and anonymized throughout the ML process(es). It is likely these issues will continue to be a concern for ML models and potentially more of a

concern than traditional epidemiological models. The reason for this is that, as mentioned, for ML models to be most effective, vast amounts of data on predictors is warranted. However, as more and more data are linked, concerns around privacy, ability to identify individuals and safeguard that data increase. The public is already very concerned with the amount of 'big data' being collected on them (e.g., recent lawsuits around Google, Facebook) and most are currently not even aware of the existing data used in most health care studies. As ML models become more mainstream, concerns around the use of this data are also expected to grow as the public becomes more informed.

Lack of ML interpretability and reporting guidelines hinder ML implementation into health systems[56,71]. The complexity of ML models and inability to easily interpret results leads to issues regarding transparency; health systems and individuals must know why ML classifiers predict the way they do. "Black boxes" are considered unacceptable by health systems[59]. Inconsistencies in reporting still contribute to the "failure to launch" of ML prediction in health systems[56]. Without measuring utility, value, and impact of ML prediction, health systems cannot stratify people within an at-risk population.

Finally, as digital technologies like ML prediction become entrenched in health systems, new human resources talented in the digital sector must become part of the health system team as new technology-oriented roles emerge[71]. Individuals need to be trained who can bridge this skills gap so that health systems can fully engage with ML technology. This is currently a major issue as often data engineers, although very skilled with data, fail to understand the nuances of health data. Conversely, people trained in health and traditional epidemiological methods fail to understand the nuances of ML. Thus, new models of training will be required to ensure all pieces required for successful ML programs are incorporated into working environments.

Finally, regulatory frameworks are required that include data governance and ethical considerations[71] in addition to ML transparency and reporting guidelines. The projects in this PhD will attempt to address some of these issues to better inform health system planners.

Summary

In summary, health systems can implement decision aid tools that fall on the ML spectrum to target interventions at high-risk groups at a population level. Benefits of ML include scalability to the population level, less human dependency, and modeling complex data. These ML classifiers can be developed on jurisdiction specific data and be continually improved with new data. Key issues impede ML implementation by health systems. The studies in this PhD program align with population health issues commonly identified by health systems and studied in the literature[16]. The ML classifiers will be trained on structured administrative data specific to Alberta, Canada and assessed using metrics and cost saving simulations that health system planners will find informative. As will be shown in subsequent chapters, using administrative health data may or may not be sufficient to satisfy the needs of a public health intervention program.

References for Chapter 2

11. 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. Alberta Machine Intelligence Institute. Machine Learning Process Lifecycle. In:2019.
16. Morgenstern JD, Buajitti E, O'Neill M, et al. Predicting population health with machine learning: a scoping review. *BMJ Open.* 2020;10(10):e037860.
17. 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.
18. Luo W, Phung D, Tran T, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res.* 2016;18(12):e323.
19. 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.
21. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *Bmj.* 2016;352:i6.
22. 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.
29. Ngiam KY, Khor W. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology.* 2019;20(5):e262-e273.
30. sonix.ai. What's the difference between artificial intelligence (AI), machine learning (ML) and natural language processing (NLP)? https://sonix.ai/articles/difference-between-artificial-intelligence-machine-learning-and-natural-language-processing. Accessed August 2022.
31. Komorowski M. Clinical management of sepsis can be improved by artificial intelligence: yes. *Intensive Care Medicine.* 2020;46(2):375-377.
32. Cognilytica. Cognitive Project Management for Artifical Intelligence Methodology. In:2020.
33. AI Enabled Care. Paper presented at: Building Collaboration for Deeper Learning and Better Care; April 29, 2021, 2021; Virtual.
34. Mooney SJ, Pejaver V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annual Review of Public Health.* 2018;39(1):95-112.

35. Canadian Institutes of Health Research. The Knowledge to Action Process. 2016; https://cihr-irsc.gc.ca/e/29418.html#6. Accessed August 2022.
36. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *JAMA.* 2018;319(13):1317-1318.
37. 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.
38. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European heart journal.* 2014;35(29):1925-1931.
39. Rose S. Machine Learning for Prediction in Electronic Health Data. *JAMA Network Open.* 2018;1(4):e181404-e181404.
40. Steyerberg EW. *Clincial Prediction Models.* Springer; 2009.
41. Michard F, Teboul JL. Predictive analytics: beyond the buzz. *Annals of Intensive Care.* 2019;9(1):46.
42. Frizzell JD, Liang L, Schulte PJ, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA cardiology.* 2017;2(2):204-209.
43. Ravaut M, Harish V, Sadeghi H, et al. Development and validation of a machine learning model using administrative health data to predict onset of type 2 diabetes. *JAMA network open.* 2021;4(5):e2111315-e2111315.
44. Xie H, McHugo G, Drake R, Sengupta A. Using discrete-time survival analysis to examine patterns of remission from substance use disorder among persons with severe mental illness. *Mental Health Services Research.* 2003;5(1):55-64.
45. Donders ART, Van Der Heijden GJ, Stijnen T, Moons KG. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology.* 2006;59(10):1087-1091.
46. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. *Journal of Clinical Epidemiology.*
47. Bhaskaran K, Smeeth L. What is the difference between missing completely at random and missing at random? *International Journal of Epidemiology.* 2014;43(4):1336-1339.
48. Greenland S, Finkle WD. A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *American Journal of Epidemiology.* 1995;142(12):1255-1264.

49. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research.* 2011;20(1):40-49.
50. 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.
51. Hu Z, Melton GB, Arsoniadis EG, Wang Y, Kwaan MR, Simon GJ. Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *Journal of Biomedical Informatics.* 2017;68:112-120.
52. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological methods.* 2002;7(2):147.
53. Sharafoddini A, Dubin JA, Maslove DM, Lee J. A new insight into missing data in intensive care unit patient profiles: Observational study. *JMIR medical informatics.* 2019;7(1):e11605.
54. Fletcher Mercaldo S, Blume JD. Missing data and prediction: the pattern submodel. *Biostatistics.* 2020;21(2):236-252.
55. Ding Y, Simonoff JS. An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research.* 2010;11(1).
56. Talhouk Aline. AI Predictive Analytics: pathways from research to the clinic. Paper presented at: AI Enabled Care: Building Collaboration for Deeper Learning and Better Care; April 2021, 2021; Michener Institute.
57. Canadian Institute for Health Information. Patient Cost Estimator: Methodology Notes and Glossary. 2022; https://www.cihi.ca/sites/default/files/document/patient-cost-estimator-methodology-notes-2021-en.pdf. Accessed May 2022, 2022.
58. Glussich A. Estimating Costs of Hospital Stays. 2016 CADTH Symposium; 2016; Ottawa, ON, Canada.
59. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA.* 2018;320(21):2199-2200.
60. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Paper presented at: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining2015.
61. McGinn TG, Guyatt GH, Wyer PC, et al. Users' Guides to the Medical LiteratureXXII: How to Use Articles About Clinical Decision Rules. *JAMA.* 2000;284(1):79-84.

62. Karthikeyan G, W. Eikelboom J. The CHADS2 score for stroke risk stratification in atrial fibrillation – friend or foe? *Thromb Haemost.* 2010;104(07):45-48.
63. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine.* 2018;178(11):1544-1547.
64. Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. *JAMA.* 2019;322(24):2377-2378.
65. Richter AN, Khoshgoftaar TM. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artificial Intelligence in Medicine.* 2018;90:1-14.
66. 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.
67. Webster LR, Webster RM. Predicting Aberrant Behaviors in Opioid-Treated Patients: Preliminary Validation of the Opioid Risk Tool. *Pain Medicine.* 2005;6(6):432-442.
68. 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.
69. Brindle PM, McConnachie A, Upton MN, Hart CL, Smith GD, Watt GC. The accuracy of the Framingham risk-score in different socioeconomic groups: a prospective study. *British Journal of General Practice.* 2005;55(520):838-845.
70. Advances in AI and Genomics: Creating a Revolution in Healthcare. May 21, 2021, 2021; Edmonton, Alberta.
71. Reznick R HK, Horsely T, Hassani MS. *Task Force Report on Artificial Intelligence and Emerging Digital Technologies.* 2020.
72. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447-453.

Chapter 3: Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada

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 emergency department visit, hospitalization 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.

Introduction

Canada is among the countries with the highest rates of opioid prescribing in the world, making prescription opioid use a key driver of the current opioid crisis[7]; a major part of the policy response to the opioid crisis focuses on endorsing safe, appropriate opioid prescribing[23,24,73]. 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[24,74,75].

Prescription monitoring programs (PMPs) have been implemented around the world, like Alberta's provincial Triplicate Prescription Program (TPP)[76] 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)[77] 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[24,37,78]; these population estimates may not be generalizable to different populations[19]. Thus, a functional gap exists in many health jurisdictions where much of the available administrative health data is not being leveraged for opioid prescription monitoring.

Supervised machine learning (ML)[11,15] is an approach that uses computer algorithms to build predictive models in the clinical setting that can make use of the large amounts of available administrative data[12,13], all within a well-defined process[14]. Supervised ML trains on labelled data to develop prediction models that are specific to different populations and, in many cases, can provide better predictive performance than traditional, population-based statistical models[13,37,79]. We identified one study[37] that applied ML techniques to predict

overdose risk in opioid patients pursuant to a prescription. In their validation sample, they found that the deep neural network (DNN) and gradient boosting machines (GBM) algorithms carried the best discrimination performance based on estimated c-statistics and that the ML approach out-performed the guideline approach in terms of risk prediction; neural networks have little interpretability and are not necessarily better at predicting outcomes when trained on structured data[60]. This study relied on c-statistics to evaluate their ML models and did not emphasize other performance metrics (e.g., positive likelihood ratios, pre and post-test probabilities) required to assess clinical utility that are recommended by medical reporting guidelines[11,17,19,20]. It also did not address the important issue of ML model interpretability[59]. Reporting informative prognostic metrics is needed to better understand the capabilities of ML classifiers if health departments and PMPs are to incorporate them into their decision-making processes.

The objective of our study was to further develop and validate ML algorithms (beyond just DNN) to predict the 30-day risk of emergency visit, hospitalization and mortality for a patient in Alberta, Canada at the time of an opioid dispensation using administrative data routinely available to health departments and PMPs and evaluate them using the above referenced reporting guidelines. We also analyzed feature importance to provide meaningful interpretations of the ML models. Comparing discrimination performance (area under the receiver operating characteristics curves), we hypothesized that the ML process would perform better than the current guideline approach for predicting risk of adverse outcomes related to opioid prescribing.

Methods

Study Design and Participants

This prognostic study used a supervised ML scheme. All patients in Alberta, Canada who received a dispensation for an opioid, were 18 years of age and older between Jan 1, 2017 and Dec 31, 2018 were eligible. Patients were excluded from all analyses if they had any previous diagnosis of cancer, received palliative interventions or were pregnant during the study period (Table 3.4) as use of opioids in these contexts is clinically different.

Government health departments and payers in many jurisdictions have systems to capture prescription histories and ICD diagnostic codes. As such, we linked various administrative health data sets available in Alberta, Canada using unique patient identifiers in order to establish a complete description of patient demographics, drug exposures and health outcomes. These databases include *1) Pharmaceutical Information Network (PIN):* PIN data includes all dispensing records from community pharmacies from all prescriber types occurring in the province outside of the hospital setting. PIN collects all drug dispensations irrespective of age or insurance status in Alberta; Anatomical Therapeutic Chemical classification (ATC) codes[80] were used to identify opioid dispensations and their respective opioid molecules (Table 3.8), 2*) Population and Vital Statistics Data (VS, Alberta Services)*: sex, age, date of birth, death date, immigration and emigration data, and underlying cause of death according to the World Health Organization algorithm using ICD codes[77], *3) Hospitalizations and Emergency Department Visits (National Ambulatory Care Reporting System [NACRS], Discharge Abstract Database [DAD]):* all services, length of stay, diagnosis (up to 25 ICD-10[77] based diagnoses). Data and coding accuracy are routinely validated both provincially and centrally via the Canadian Institute for Health Information, and 4*) Physician Visits/Claims (Alberta Health):* all claims from all settings (e.g., outpatient, office visits, emergency departments, inpatient) with associated date of service, ICD code, procedure and billing information.

This study followed the TRIPOD and STARD reporting guidelines[22,68,81] and received ethics approval from the University of Alberta ethics board (Pro00083807_AME1).

Measures and Outcome

ML models were trained on a labelled dataset in which the observation/analysis unit was an opioid dispensation. Every opioid dispensation, not just the incident one, was used as a potential instance to predict the risk of our outcome. The primary outcome was a composite of a drug-related emergency department (ED) visit, hospitalization or mortality within 30 days of an opioid dispensation based on ICD-10 codes used by others and identified from DAD, NACRS and Vital Statistics (T40, F55, F10-19; Table 3.5)[23,37,82].

We anticipated that our defined outcome would be a rare event, leading to a class imbalanced dataset[83]. To address this, we relied on specifying balanced class weightage for supporting algorithms; other approaches were deemed not suitable (e.g., oversampling using randomly repeating minority class); under sampling (sub-sampling within the majority class) resulted in changes in outcome prevalence. Class weightage is a commonly used method[84] to address class imbalance along with over and under-sampling approaches. However, oversampling, which involves generating new opioid dispensations from the original data distribution and is prone to introducing bias, is difficult due to the categorical nature of the data and beyond the scope of this study. With under-sampling, which takes samples from the majority class (in this case, no 30-day event after dispensation), we would not be able to use all of the information provided by the data in instances with no outcome. Hence, we decided to use the class weightage method which does not alter the data distribution. Instead, the learning process is adjusted in a way that increases the importance of the positive class (instances that led to a 30-day event)[85].

Predictor Candidates for ML Models

Predictor variables in our ML models included those that were informed by the literature[24,37,73] and those directly obtained from the data sets. These included features based on demographics (age, sex, income using Forward Sortation index from postal codes[86]), co-morbidity history using ICD-based Elixhauser score categories[87], health care utilization (number of unique providers, number of hospital and emergency department visits), and drug utilization (level 3 ATC codes[80], oral morphine equivalents[88], 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 (Figure 3.5); 30 days was used to reflect the immediate nature of the risk and 5 years to fully capture co-morbidities. This approach aligns with how health providers would assess patients using the entire history of co-morbidities and then the more immediate factors in deciding on the need for a therapeutic as well as risk in patients. We performed experiments to identify the

features and data sets that contributed most to predicting the outcomes with a view to minimizing the potential future data requirements for health departments and PMPs.

Statistical Analyses and Machine-Learning Prediction Evaluation

We randomly divided the patients in the 2017 portion of our study cohort into training (70%) and validation (30%) sets[11] by patients and opioid dispensations such that no patients in the training set were in the validation set. Baseline characteristics and event rates were compared in the training vs validation group, and between those who experienced the outcome and those who did not using chi-squared tests of independence. As well, we used all the 2018 data as another independent validation set.

We trained commonly used[11,50] ML algorithms (Appendix in Chapter 3) and further tuned out-of-box models using 5-fold cross validation on the training data to address model overfitting[11,39]. As is common in ML validation studies[11,37], we reported model discrimination performance (i.e. how well a model differentiates those at higher risk from those at lower risk)[19] using area under the receiver operating characteristic curve (AUROC; c-statistic). We then stratified the two ML models with the highest c-statistics into percentile categories (deciles) according to absolute risk of our outcome, as was done in previous studies[37,66]. We also plotted AUROC[19] and precision-recall curves (PRCs)[89].

Because discrimination alone is insufficient to assess ML model prediction capability, we assessed a second necessary property, namely, calibration (i.e., how similar the predicted absolute risk is to the observed risk across different risk strata)[19,90]. Using the two ML models with the highest discrimination performance, we assessed calibration performance on the 2018 data by plotting observed (fraction of positives) vs predicted risk (mean predicted value). Using these same two ML classifiers, we analyzed the top 0.1, 1, 5, and 10 percentiles of predicted risk by the number of true and false positives, positive likelihood ratios (PLR)[20], positive predictive values (PPV), post-test probabilities, and number needed to screen. We also performed a simulation of daily data uploads for 2018 Quarter 1 to view the predictive capabilities if a ML risk predictor were to be deployed into a monitoring workflow.

For the XGBoost and logistic regression classifiers, we reported feature importance[50] and plotted PRCs that compared all dispenses to those within the top 10 percentiles of estimated risk. As well, for the XGBoost classifier, we described feature importance on model outcome using SHAP values[91,92] to add an additional layer of interpretability.

Finally, we compared ML risk prediction (the two ML models with highest discrimination performance) to current guideline approaches as others have[37], using the 2019 Centers for Medicare & Medicaid Services (CMS) opioid safety measures[93] and the 2017 Canadian Opioid Prescribing Guideline[24]. This was done by using the guidelines as "rules" when coding for the 30-day risk of event at the time of each opioid dispensation on the entire 2018 validation set. We also compared the discrimination performance of different logistic regression classifier models using various combinations of features derived from their respective databases: **1)** demographic and drug/health utilization features from PIN and 2**)** co-morbidity features derived from DAD, NACRS and Claims.

All analyses were done using Python (v. 3.6.8,), SciKit Learn[94] (v. 0.23.2) SHAP[92] (v. 0.35), XGBoost (v. 0.90)[95], Pandas (v. 1.0.5)[96] and H20 Driverless AI (version 1.9).

Patient and Public Involvement

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

Results

Patient Characteristics and Predictors

We identified 392,979 patients with at least one opioid dispensation in 2017 (Figure 3.1). 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 (Table 3.6) while characteristics were similar between the

training and validation sets (Table 3.7). 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 (Table 3.8). When using all the databases, the total number of features was 283 and 34 when considering only co-morbidities.

Machine-Learning Prediction Performance

Using the 2017 validation set, AUROCs for the XGBoost and logistic regression classifiers had the highest discrimination performance at 0.87, while the neural network classifier had lower performance at 0.80 (Table 3.9).

Discrimination performance was similar for the 2018 validation set (n=393,023; Table 3.9). XGBoost and logistic regression had the highest estimated AUROCs and area under PRCs while the neural network classifier was lower (Figure 3.2A, 3.2B). As expected, precision-recall curves indicate stronger predictive performance in opioid dispensations at higher predicted risk percentiles (Figure 3.2C, 3.2D).

In the 2018 validation set, although discrimination performance was similar (0.88), individual feature importance was different between the logistic regression and XGBoost classifiers, with logistic regression feature importance more reliant on co-morbidity data from DAD, NACRS and Claims while XGBoost relied more on drug utilization data from PIN (Figure 3.6). With the XGBoost classifier, history of drug abuse, alcoholism, and prior hospitalization/emergency visit carried the highest importance for predicting the study outcome (Figure 3.7A) where the presence of these features in a patient suggested a strong prediction towards having the defined outcome (Figure 3.7B and 3.7C).

Calibration

When considering dispensations predicted to be in the highest percentiles of risk, the top 5-percentile captured 42% of all outcomes using the XGBoost and logistic regression classifiers (Table 3.1). Also, as the predicted risk percentiles get higher (top 10 percentile to top 0.1 percentile), so too do the corresponding PPVs with the top 0.1 percentile associated with a

PPV of 33% for the XGBoost classifier. As well, lower categories of risk percentiles were associated with lower outcomes (Figure 3.3, Figure 3.8). When we simulated a monitoring workflow scenario with daily data uploads, a similar pattern was illustrated where the dispensations predicted to be higher risk had higher event rates (Figure 3.4).

After using the XGBoost and logistic regression classifiers to identify the dispensations in the highest predicted risk percentiles, the pre-test probability of the outcome (1.6%) was transformed into higher post-test probabilities, with higher probabilities in the riskier percentiles (Table 3.1). The number needed to screen also decreased as predicted risk increased (Table 3.1).

Comparing discrimination performance, ML risk prediction outperformed the current guideline approaches when using various combinations of guideline recommendations (Table 3.2). In many of the guideline scenarios, the estimated AUROCs were close to the 0.5 mark. When we estimated the discrimination performance of the logistic regression classifier based on database source, using all databases produced an AUROC of 0.88. Reducing the database source to only DAD, NACRS, Claims (co-morbidities only) resulted in an AUROC of 0.85, while PIN (prescription history) only was 0.78 (Table 3.3).

Discussion

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

Since the prevalence of our defined outcome is relatively low in the general population, PPVs would also be expectedly low. However, estimated PPVs increased when we considered

higher risk dispensations, as is expected since PPV is related to event prevalence. This is important because different users of a risk predictor will require different predictive capabilities. Similarly, our estimates of positive likelihood ratios and associated post-test probabilities also increased in dispensations with higher predicted risk indicating the strong predictive capabilities of the XGBoost and logistic regression classifiers; likelihood ratios >10 generate conclusive changes from pre-test to post-test probabilities[20].

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

We found only one study that used ML approaches to quantify the absolute risk of an event pursuant to an opioid dispensation[37]. Their methodology used rolling 3-month windows for estimating risk and ML model training while we used historic records to estimate 30-day risk. Differences in study population and feature selection may explain why their highest performing ML model was deep learning (neural network classifier) and ours was not. Nevertheless, we were able to replicate their predictive performance using our ML approach as we both showed that ML approaches have higher predictive capabilities than guideline approaches. Both of our studies used predicted percentile risk estimates to identify high risk dispensations and were able to do so with strong discrimination and calibration performance. Furthermore, we emphasized prognostic metrics which are more informative to assess the clinical utility of ML classifiers using pre- and post-test probabilities, something not done in other studies and recommended in medical guidelines[20]. This major aspect of our study, not done previously, is important because any ML classifier that does not increase prognostic information compared to baseline cannot be incorporated into decision making for the purpose

of intervening on higher risk instead of lower risk patients. Indeed, another study we found describes how identifying cases in higher predicted risk percentiles using ML methods can be deployed in hospital settings for the purpose of targeted interventions[66] upon discharge, however the effect on outcomes is still to be determined.

The limitations of our study are similar to other ML studies[37] and need to be addressed when considering deployment of ML risk predictors. Our training dataset was not able to account for non-prescription opioid consumption and the risk associated with non-prescription use, both of which are substantial contributors to overall risk[23]. Regarding our analysis, we assumed that all dispensations were independent events; future research in this area should focus on employing ML methods using correlated data. As with all ML projects, our models were trained using Alberta data and might not be generalizable to other populations, or to specific populations within Alberta. However, one of the benefits of the ML process is that models can be retrained or similar methods could be used to develop new models to accommodate different populations.

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

Table 3.1. 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 3.2. Discrimination performance of guideline approach using the 2018 validation set. Guideline approaches were adapted from the 2017 Canadian Opioid Prescribing Guideline and 2019 Centers for Medicare & Medicaid Services (CMS) opioid safety measures and compared to logistic regression and XGBoost classifiers (each with an estimated area under the receiver operating characteristic curve of 0.88). These guidelines were used as rules to predict the 30-day risk of event at the time of opioid dispensation.

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

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

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

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

Table 3.3. Discrimination performance based on database source using area under the receiver operating characteristic curve (AUROC) for the logistic regression classifier on the 2018 validation set.

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

Note: drug utilization includes features describing oral morphine equivalents[88], 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 3.1. 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

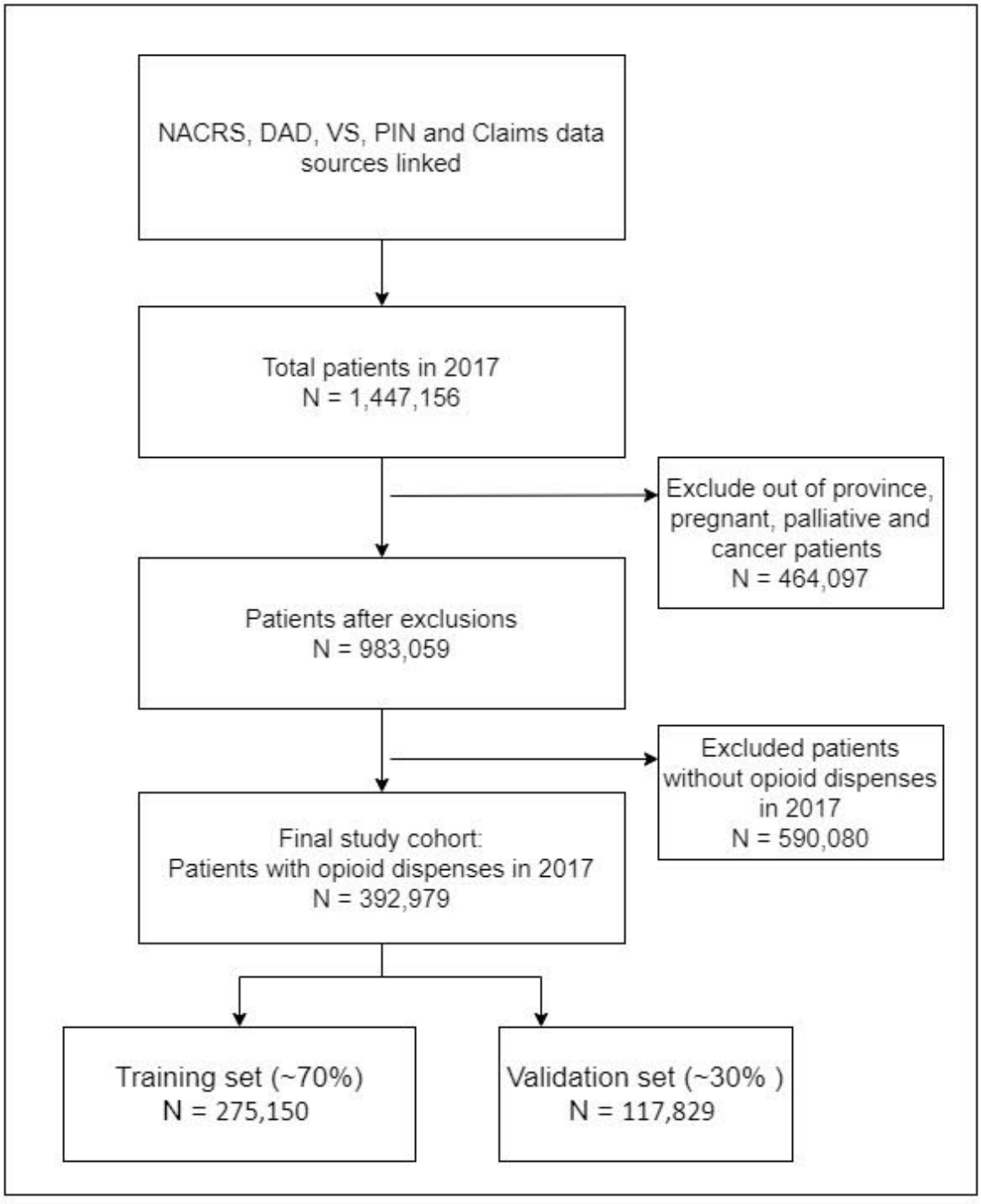
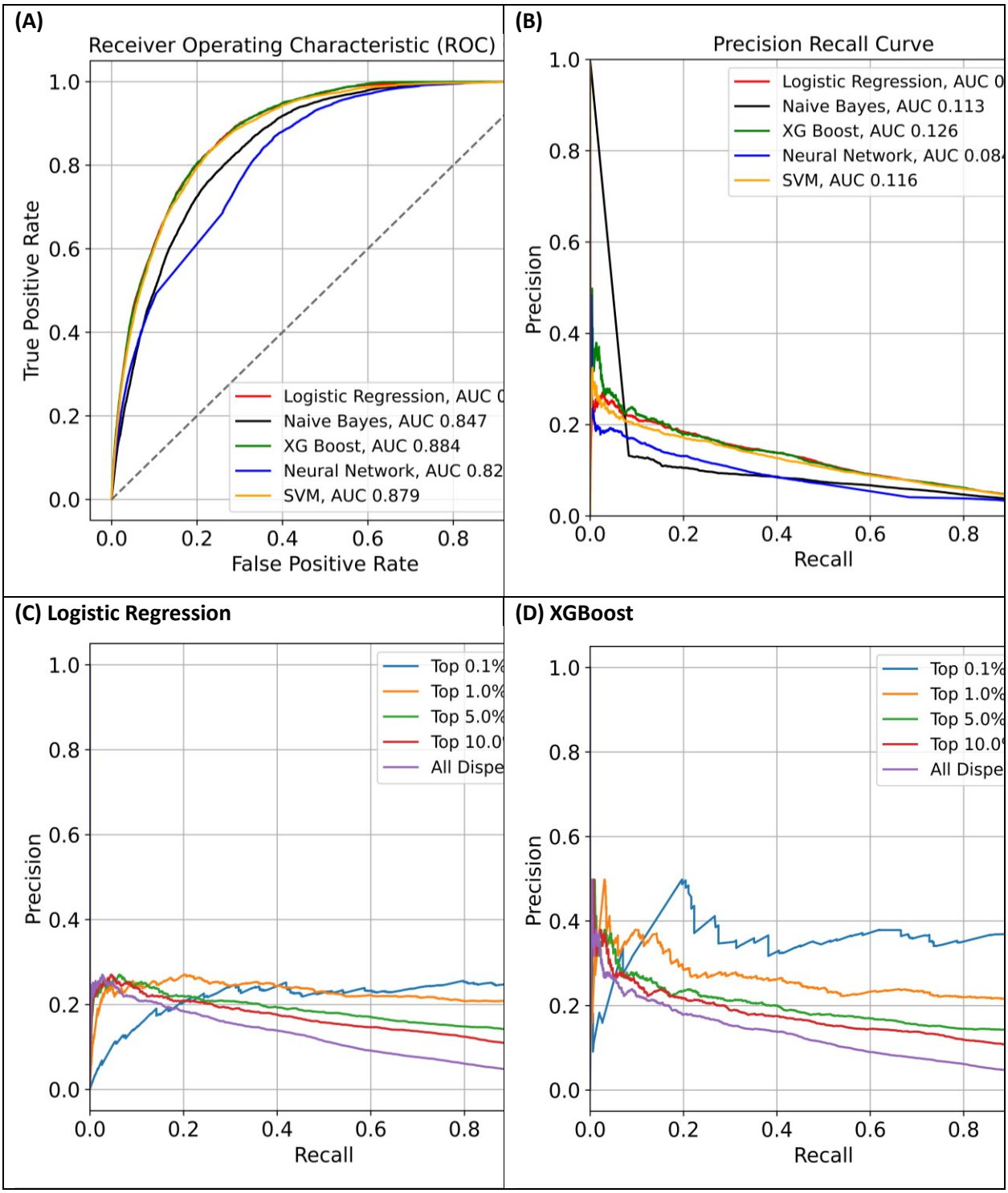


Figure 3.2. Area under the receiver operating characteristic curves (A) and precision-recall curves (B) for all dispensations using logistic regression (L1), neural network, support vector machine (SVM), XGBoost and Naïve-Bayes; precision-recall curves for higher risk dispensations according to predicted risk percentile categories for logistic regression (C) and XGBoost (D) using the 2018 validation set.



AUC: area under the curve

Figure 3.3. Calibration curve plotting observed vs. quantiles (deciles) of estimated risk for the XGBoost classifier using the 2018 validation dataset. Most counts (dispensations) were predicted to be lower risk.

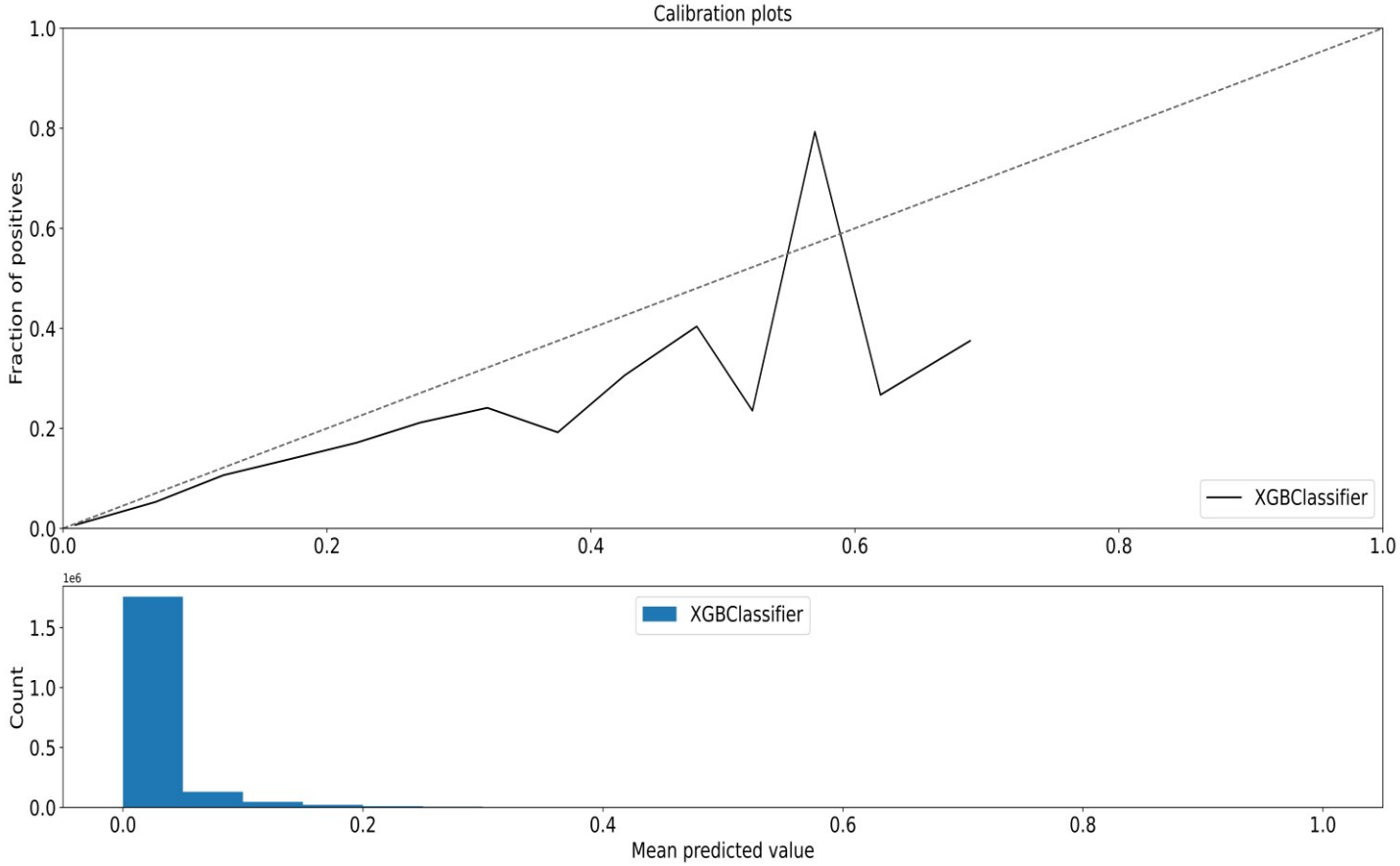
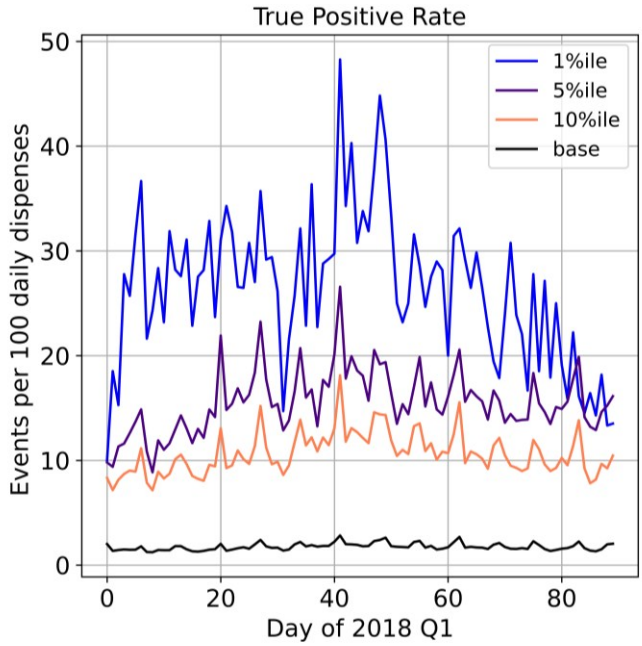
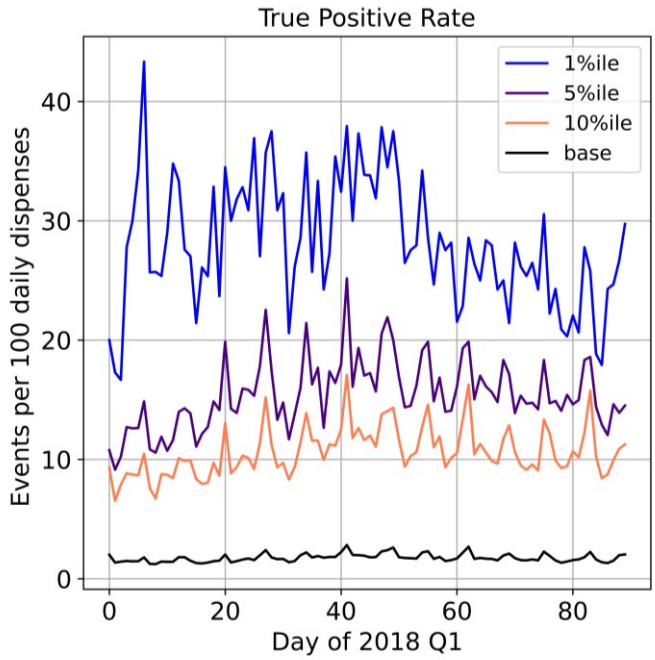


Figure 3.4. Simulation of a clinical workflow with daily uploads and events per 100 daily dispenses by risk percentiles using 2018 Quarter 1 (Q1) data for logistic regression (A) and XGBoost (B) classifiers.

(A) Logistic Regression (L1)



(B) XGBoost



Appendix to Chapter 3

Introduction

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

Logistic Regression

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

Ridge Classifier

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

Random Forest

Random forest is a tree ensemble learning algorithm that has wide applicability in many domains[1]. Random forest is a nonlinear learning algorithm, which arrives at nonlinear decision boundaries by independently combining multiple decision trees. Each individual decision tree in the forest can be grown independently of each other on a subset of the training data. Random

forests are mainly sensitive to the number of trees, the depth of a tree and the number of covariates randomly chosen to split at each node[1]. These hyper-parameters can be tuned to find the best configuration of every dataset. Random Forests, in general, are less prone to overfit since they always grow individual trees on a subset of the training data[1]. At prediction time, the decision of each tree is aggregated to compute the final prediction.

Neural Networks (NN)

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

Boosted Learning Algorithms

Boosting is a process to ensemble multiple base learning algorithms to arrive at better overall performance than any individual base learner[1]. In contrast to independently building multiple models from the subsets of the data, boosting re-weights the training data every time a model is learned for future models. This weighting happens to give more preference to currently misclassified data points in the next round compared to the correctly classified data points. Therefore future learners try to do better on the misclassified data points leading to a collection

base learners having a better-combined prediction. This process is sequential so each base learner is dependent on the output of the previously trained model (it is worthy to note XGBoost provides a parallel tree boosting alternative). In our work, we have experimented with several boosting meta-learning algorithms such as XGBoost[7], AdaBoost[5] and GBM[5]. XGBoost uses a variant of trees as the base learner whereas AdaBoost (from Sci-kit learn) can use many ML algorithms as base learners. GBM uses logistic regression by default as the base learner. We used all 3 types of boosting with tuned hyperparameters for comparison.

Naive Bayes

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

Support Vector Machines (SVM)

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

References for Appendix

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).

Table 3.4. 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

Table 3.5. 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

Table 3.6. Baseline characteristics of study patients (n=392,979). Co-morbidities were determined using Elixhauser criteria. All p-values in the chi^2 test of independence were <0.001 unless otherwise indicated.

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

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

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

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

Table 3.7. Characteristics of study participants between training and validation groups using 2017 data.

Characteristic	Number in training group N=275,150~	Percent	Number in validation group N=117,829~	Percent
Age:				
Mean (SD)	48.3 (16)	--	48.2 (16)	--
18-45	114356	41.5	49909	42.3
45-65	111859	40.7	47132	40.0
>65	48935	17.8	20788	17.6
Male	138603	48.5	59339	48.4
Female	136545	47.8	58490	47.7
Alcohol Disorder	46792	16.4	20199	16.5
Arrhythmia	63637	22.3	27201	22.2
Blood Loss Anemia	839	0.3	336	0.3
Congestive Heart Failure	13320	4.7	5694	4.6
Coagulopathy	5697	2.0	2393	2.0
Deficiency Anemia	24096	8.4	10179	8.3
Depression	112080	39.2	47628	38.9
Diabetes**	45131	15.8	19144	15.6
Substance Abuse Disorder	52609	18.4	22713	18.5
Fluid Disorder	30272	10.6	12780	10.4
Hypertension**	98546	34.5	41840	34.1
Hypothyroidism	31908	11.2	13666	11.2
Injury*	137423	48.1	58865	48.0
Liver Disorder	15252	5.3	6567	5.4
Neurologic Disorder	161706	56.5	69341	56.6
Obesity	44607	15.6	18882	15.4
Poisoning*	12503	4.4	5293	4.3
Psychoses	25422	8.9	10860	8.9
Renal Disorder	11403	4.0	4817	3.9
Rheumatoid Conditions	78268	27.4	33420	27.3
HIV Infection	774	0.3	336	0.3
Paralysis	2717	1.0	1176	1.0
Peptic Ulcer Disease	8239	2.9	3533	2.9
Pulmonary Circulation Disorder	6771	2.4	2877	2.3
Chronic Pulmonary Disease	72265	25.3	30949	25.3
Peripheral Vascular Disease	10228	3.6	4278	3.5
Valvular Disease	5111	1.8	2215	1.8
Weight Loss	11477	4.0	4790	3.9

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

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

Table 3.8. 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

Table 3.9. 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

Figure 3.5. Schematic of study design and feature generation

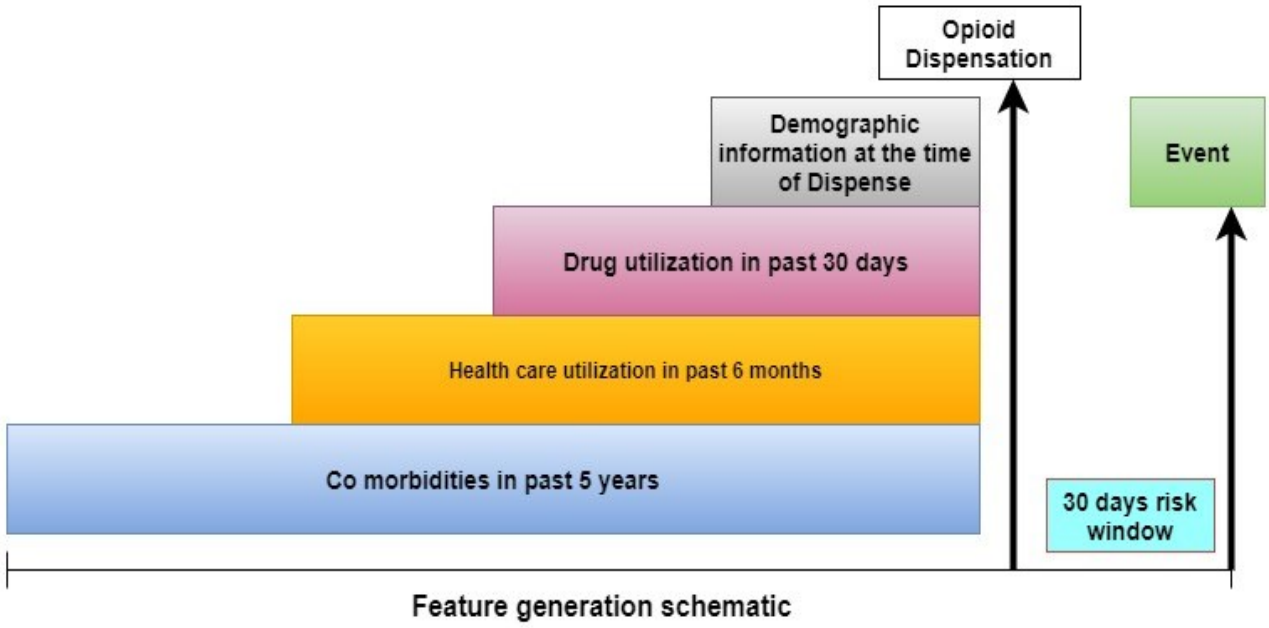


Figure 3.6. Feature importance from logistic regression and tree-based XGBoost classifiers using the 2018 validation set. The logistic regression classifier relied more on co-morbidity data from DAD, NACRS, and Claims databases; XGBoost classifier relied more on data from the PIN database. AUROCs for 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

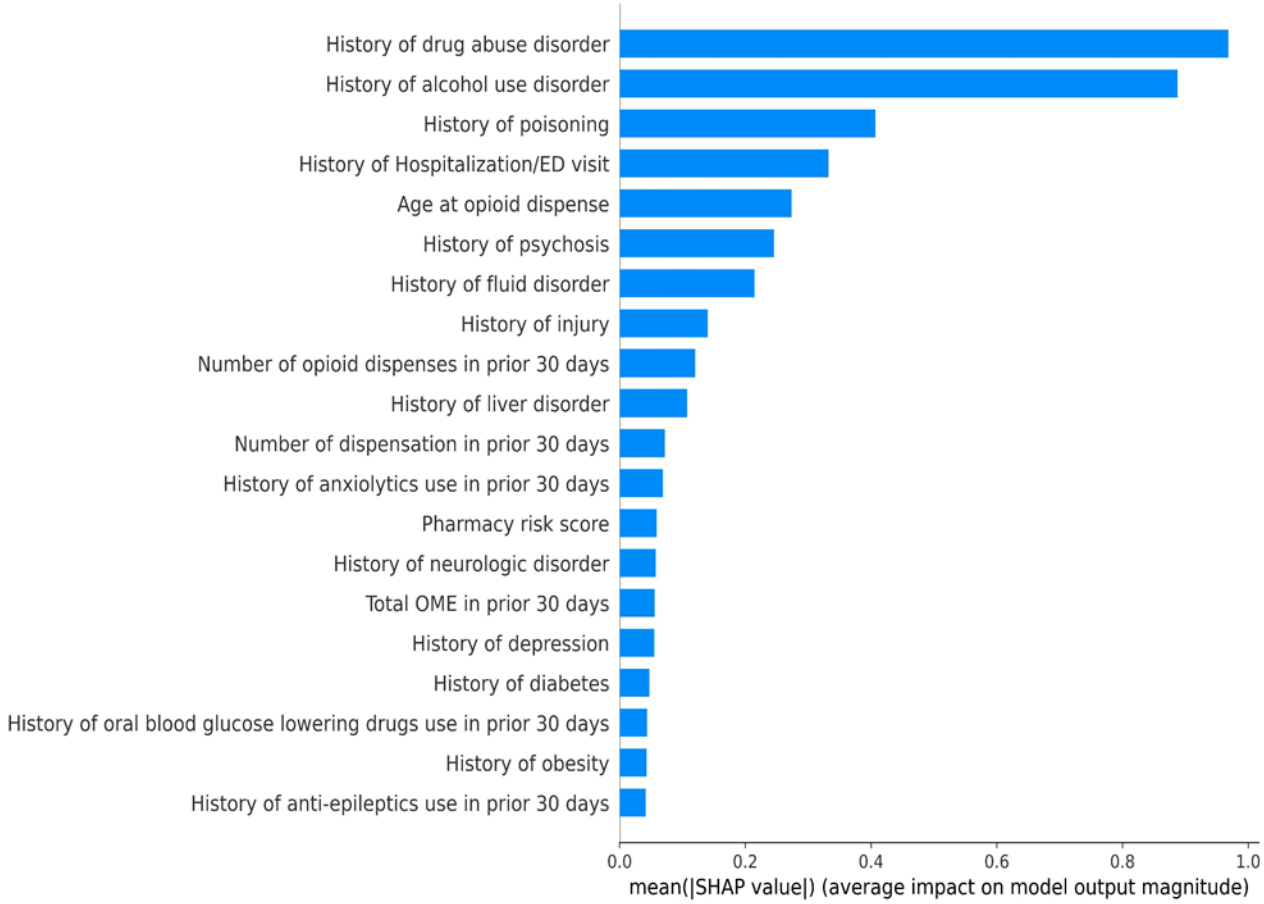
Note: Logistic regression used L1 (lasso) parameter regularization; BZD- benzodiazepine; OME- oral morphine equivalents; ED: emergency department

*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

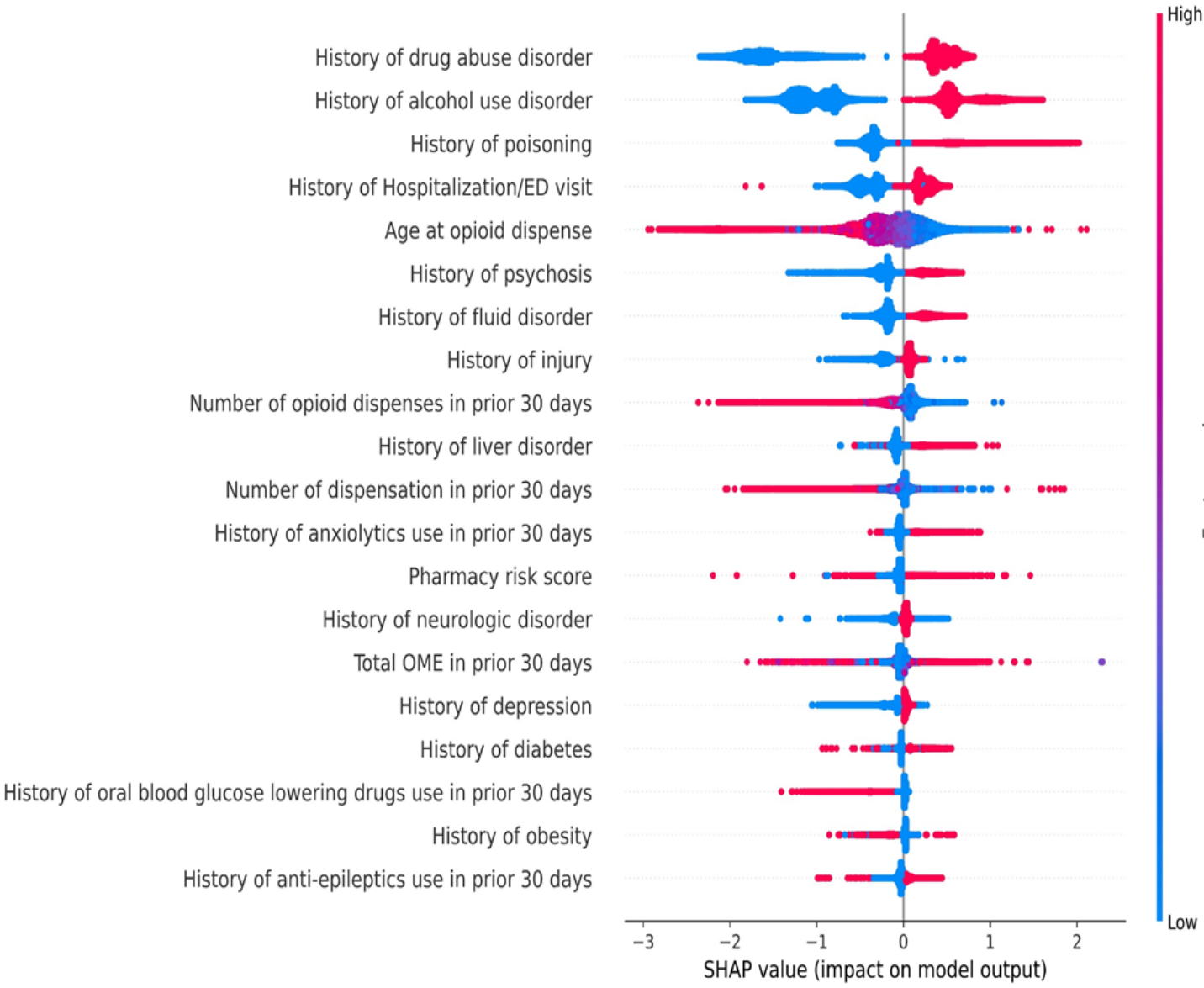
Figure 3.7. SHAP values and feature impact of the XGBoost classifier using the 2018 validation set to describe "associations" between features and the outcome. Features with the most impact on the model with drug abuse ranked highest (A); tornado plot illustrating feature impact (B); explaining the prediction of study outcome based on predictor values for 4 patients using SHAP values(C).

(A)



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

(B)



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

(C)

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

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

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

Figure 3.7C

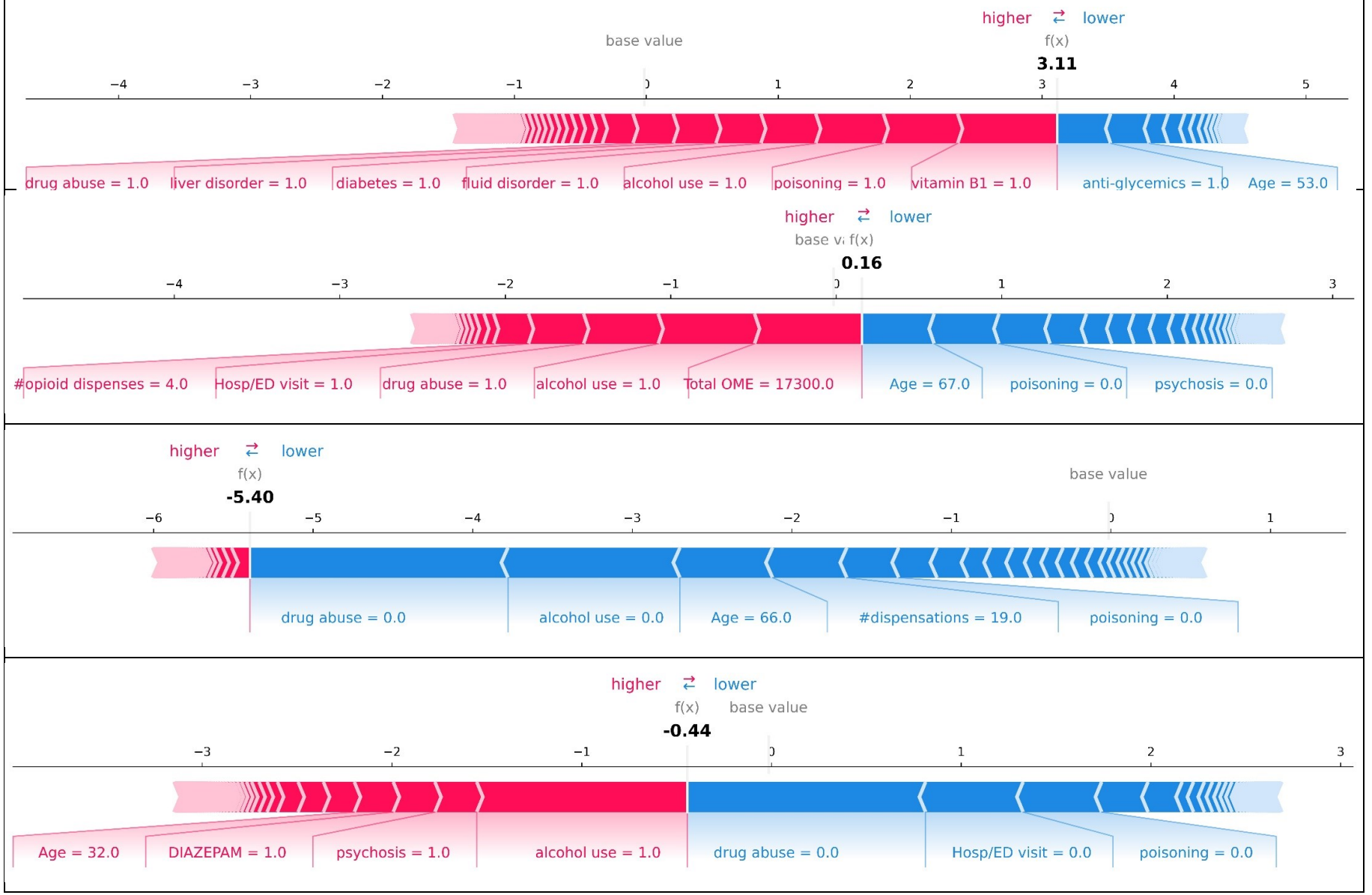
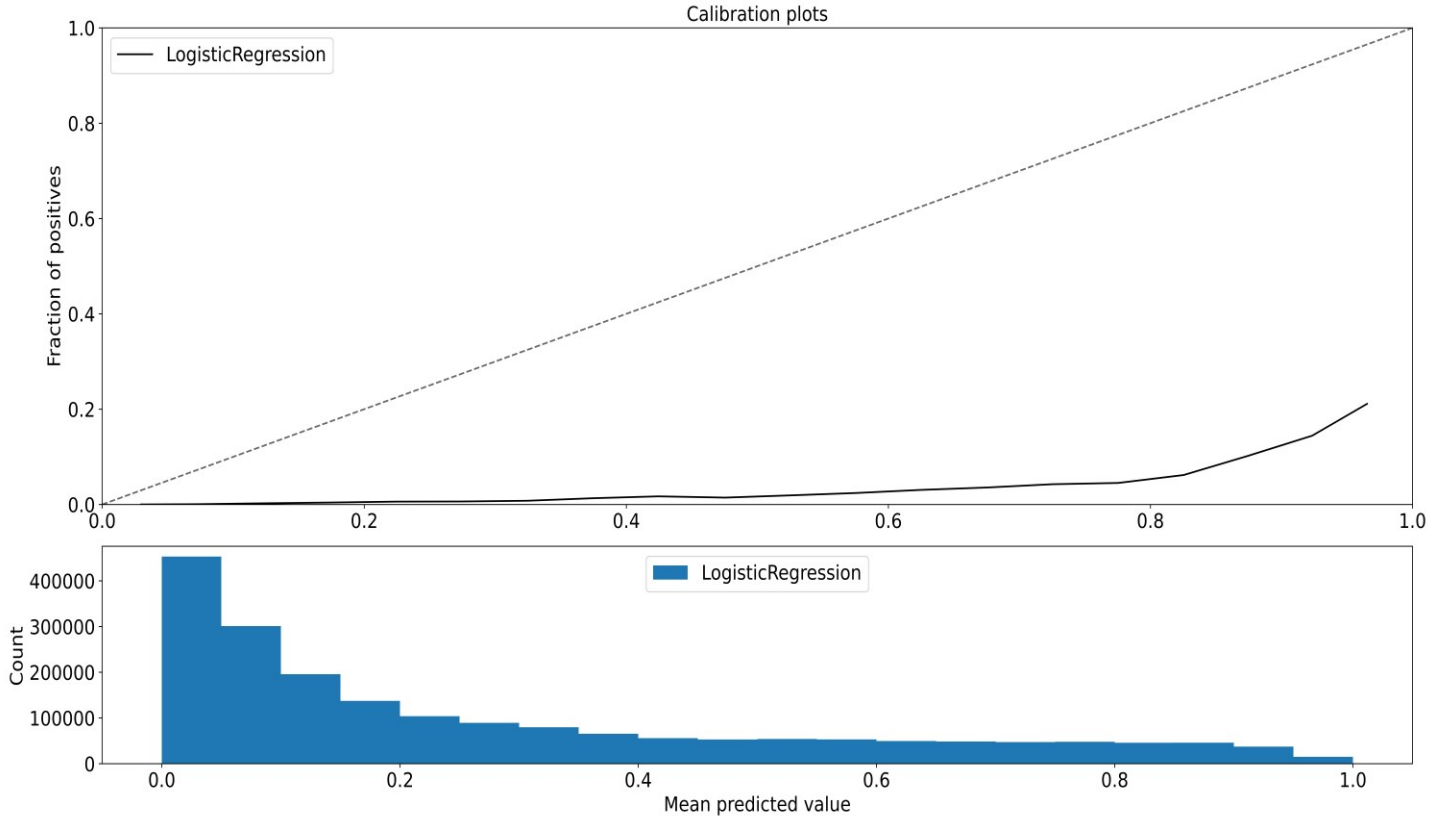

higher ⇄ lower
base value
f(x)
3.11
drug abuse = 1.0
liver disorder = 1.0
diabetes = 1.0
fluid disorder = 1.0
alcohol use = 1.0
poisoning = 1.0
vitamin B1 = 1.0
anti-glycemics = 1.0
Age = 53.0
higher ⇄ lower
f(x)
0.16
#opioid dispenses = 4.0
Hosp/ED visit = 1.0
drug abuse = 1.0
alcohol use = 1.0
Total OME = 17300.0
Age = 67.0
poisoning = 0.0
psychosis = 0.0
higher ⇄ lower
f(x)
-5.40
base value
drug abuse = 0.0
alcohol use = 0.0
Age = 66.0
#dispensations = 19.0
poisoning = 0.0
higher ⇄ lower
f(x)
base value
-0.44
Age = 32.0
DIAZEPAM = 1.0
psychosis = 1.0
alcohol use = 1.0
drug abuse = 0.0
Hosp/ED visit = 0.0
poisoning = 0.0

Figure 3.8. Calibration curve plotting observed vs. quantiles of estimated risk for the logistic regression (L1) classifier using the 2018 validation dataset. Most counts (dispensations) were predicted to be lower risk.



References for Chapter 3.

7. 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.
11. 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.
12. Bastanlar Y, Ozuysal M. Introduction to machine learning. *Methods in molecular biology (Clifton, NJ).* 2014;1107:105-128.
13. 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.
14. Alberta Machine Intelligence Institute. Machine Learning Process Lifecycle. In:2019.
15. Shah NH, Milstein A, Bagley P, Steven C. Making Machine Learning Models Clinically Useful. *JAMA.* 2019;322(14):1351-1352.
17. 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.
19. 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.
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.
22. 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.
23. 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.
24. 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.
37. 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.

39. Rose S. Machine Learning for Prediction in Electronic Health Data. *JAMA Network Open.* 2018;1(4):e181404-e181404.

50. 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.

59. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA.* 2018;320(21):2199-2200.

60. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Paper presented at: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining2015.

66. 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.

68. 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.

73. Dowell D. CDC guideline for prescribing opioids for chronic pain. 2016.

74. ismp Canada. Essential Clinical Skills for Opioid Prescribers. 2017; https://www.ismp-canada.org/download/OpioidStewardship/Opioid-Prescribing-Skills.pdf. Accessed Nov 2018.

75. Centre for Effective Practice. Management of Chronic Non Cancer Pain. 2017; thewellhealth.ca/cncp.

76. College of Physicians and Surgeons of Alberta. TPP Alberta – OME and DDD Conversion Factors. 2020; http://www.cpsa.ca/tpp/. Accessed Jun 2020.

77. World health Organization. Classification of Diseases (ICD). 2019; https://www.who.int/classifications/icd/icdonlineversions/en/. Accessed Jun 2020.

78. Gomes T, Mamdani MM, Dhalla IA, Paterson JM, Juurlink DN. Opioid Dose and Drug-Related Mortality in Patients With Nonmalignant PainOpioid Dose and Drug-related Mortality. *JAMA Internal Medicine.* 2011;171(7):686-691.

79. 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.

80. World Health Organization. International language for drug utilization research, ATC/DDD. 2020; https://www.whocc.no/. Accessed Jun 2020, 2020.

81. 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.
82. 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.
83. Brownlee J. A Gentle Introduction to Imbalanced Classification. 2020; https://machinelearningmastery.com/what-is-imbalanced-classification/. Accessed Jan 2021.
84. King G, Zeng L. Logistic regression in rare events data. *Political analysis.* 2001;9(2):137-163.
85. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. *Journal of Big Data.* 2019;6(1):1-54.
86. 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.
87. 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.
88. 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.
89. 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.
90. Shah ND, Steyerberg EW, Kent DM. Big Data and Predictive Analytics: Recalibrating Expectations. *JAMA.* 2018;320(1):27-28.
91. Molnar C. *Interpretable machine learning. A Guide for Making Black Box Models Explainable.* 2019.
92. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Paper presented at: Advances in neural information processing systems2017.
93. 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.
94. 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.
95. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Paper presented at: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining2016.
96. The pandas development team. pandas-dev/pandas: Pandas. 2020; https://doi.org/10.5281/zenodo.3509134, Jan 2021.

Chapter 4: Predicting 30-day readmissions in patients with heart failure using administrative data: a machine learning approach

Aims: To develop machine-learning (ML) models trained on administrative data which predict risk of readmission in heart failure (HF) patients; evaluate and compare the ML model with the currently used LaCE score using clinically informative metrics.

Methods and Results: This prognostic study was conducted in Alberta, Canada on 9,845 patients with confirmed HF admitted to hospital between 2012-2019. The outcome was unplanned all-cause hospital readmission within 30-days of discharge. 80% of the data was used for ML model development and 20% for independent validation. We reported, using the validation set, c-statistics (AUROCs)and performance metrics (likelihood ratio [LR], positive predictive values [PPV]) for the XGBoost model and a modified LaCE score within their respective predictive thresholds. Boosted tree-based classifiers had higher AUROCs (0.65 for XGBoost) compared to others (0.58 for Neural Network) and 0.57 for the modified LaCE. Within the predicted threshold range of the XGBoost classifier, the positive LR was 1.00 at the low end of predicted risk and 6.12 at the high end, resulting in a PPV (post-test probability) range of 21-62%; the pre-test probability of readmission was 20.9% using prevalence. The corresponding positive LRs and PPVs across LaCE score thresholds were 1.00-1.20 and 21-24%, respectively.

Conclusion: Despite predicting readmissions better than the LaCE, even the best ML model trained on administrative health data (XGBoost) did not provide substantially informative prediction performance as it only generated a moderate shift from pre to post-test probability. Health systems wishing to deploy such a tool should consider training ML models with additional data. Adding other techniques like Natural Language Processing, along with ML, to use other clinical information (like chart notes) might improve prediction performance.

Keywords: heart failure, machine learning, risk prediction, readmissions, administrative data

Introduction

Despite advances in diagnosis and treatment guidelines[97], patients with heart failure (HF) have among the highest rates of unplanned 30-day hospital readmissions in Canada and the US[4,5]. Readmission rates range from 20-27% in North America[4,9] while costs to health payers are in the billions of dollars[9]. The burden of unplanned readmissions on patients, family members and health systems has resulted in growing attention to this issue, especially since at least some readmissions are potentially avoidable[25,98].

Given these potentially preventable costs, there is substantial interest from health payers[99] and health systems (e.g. transitional care teams)[100] to predict unplanned readmissions in the HF population. Thus, there is a need for prediction tools which can accommodate the heterogeneous nature of HF patients to predict readmissions.

Currently, the LACE[101] and newer LaCE[102] scores use administrative health data to assess 30-day readmission risk in patients with HF, albeit, with insufficient accuracy[102-104]. All versions of the LACE score were developed using parametric regression resulting in a risk score and not individual probabilities to predict readmissions. Machine Learning (ML) methods are also being used to train models that can predict readmissions in HF populations using the growing availability of more granular electronic medical record data[11,15,42,105]; ML is an alternative approach for clinical prediction which produces absolute probabilities at the individual level and not a risk score based on parametric regression. However, the performance of these ML classifiers is doubtful because of modest discrimination performance and the surprising paucity[106] of reporting performance metrics recommended by medical reporting guidelines[11,17,19,20,22]. ML models reporting informative prognostic metrics are needed to better understand the capabilities of this methodology if health systems are to incorporate them into their decision-making process and work flows. Nevertheless, a recent meta-analysis done by Shin et al. reported ML methods had better discrimination than conventional statistical approaches and that more informative metrics should be included with ML reporting[107].

The objective of this prognostic study is to develop and validate ML models to predict the risk of unplanned, all-cause 30-day readmissions after discharge in a previously defined HF

cohort using only administrative health data. We compared the ML classifiers with the LaCE score, a non-ML prediction tool, using clinically informative metrics specified in reporting guidelines[11,17,19,20,22]. We plan to improve prediction performance by using newer ML algorithms (e.g., XGBoost) and to use population based administrative data from Alberta to highlight the Alberta experience; others have not done this to our knowledge.

Methods

Study Design, Setting and Participants

This prognostic study used a supervised ML scheme[11] which trained ML models on hospitalizations in Alberta, Canada between April 2012 – March 2019. Our sample of patients came from a previously derived HF cohort[108] who had a cardiologist-confirmed clinical diagnosis of HF and at least 2 echocardiograms (n=10,641) in Alberta, Canada. We further restricted this sample by excluding 113 patients with no recorded hospitalizations (as a readmission was not possible), 221 who died during hospitalization, and 23 whose records were not fully captured within the administrative data (Figure 4.1).

Data Sources

We linked our final sample to administrative health databases maintained by Alberta Health Services using anonymized patient identifiers. These include *1) Pharmaceutical Information Network (PIN):* data on all dispensing records from community pharmacies irrespective of coverage status and according to the guidelines from the Alberta College Pharmacy[109], *2) Population and Vital Statistics Data (VS, Alberta Services)*: sex, age, date of birth, death date, immigration and emigration data within the province, and underlying cause of death according to the World Health Organization algorithm using ICD codes (International Statistical Classification of Diseases and Related Health Problems[77]), *3) Hospitalizations and Emergency Department Visits (NACRS [National Ambulatory Care Reporting System], DAD [Discharge Abstract Database]):* all services, length of stay, diagnosis (up to 25 ICD-10 based diagnoses), discharge dispositions (e.g., transfers, discharges). Data and coding accuracy are routinely validated both provincially and centrally via the Canadian Institute for Health

Information[110], 4*) Physician Visits/Claims (Alberta Health):* date of service, up to three ICD codes associated with the claim, procedure and billing information, and *5) Provincial Laboratory, Alberta Health Services:* all laboratory services conducted within the hospital or community.

The above-mentioned data sets span the time period from October 2000 to March 2019.

These linked databases represent a labeled data set used to train ML models and calculate the LaCE score for each instance. We used the LaCE score, which excludes admission acuity but includes age plus length of index hospital stay, Charlson comorbidity score, and emergency department use in the prior 6 months, because it has been shown to have better discrimination performance than the LACE[101,102] and is more accurate than the LACE+[111] in predicting readmissions in Alberta. While the LaCE score was developed at the patient level, its use is intended to assess readmission risk at the time of discharge, whether it be a repeat readmission or not. Furthermore, both readmission or death risk can be assessed by the LaCE and to be consistent with discharge units in Alberta, Canada and other studies, we are only considering readmission risk[66,112]. Thus, our dataset will be organized in a manner similar to hospital discharge units to allow the comparison between the LaCE score and ML models.

Measures and Outcome

Our follow-up and predictions started after the first all-cause hospitalization, the "reference hospitalization", following patients' first echocardiogram (or the concurrent hospitalization if the echocardiogram occurred while in hospital) which led to a discharge to home (with or without supportive services) or continuing care. Hospitalizations which led to death or transfers were not included as reference hospitalizations. In instances of transfers, the final hospitalization which led to discharge was used as the reference hospitalization, however, all hospitalizations in the series of transfers leading up to the reference hospitalization were used to calculate length of stay, something not done in other HF ML studies[42,113]. The unit of analysis used to train the ML models was hospital admission. Others performed their analysis at the patient level[42] or at a level not identified[114] with unremarkable prediction performance. Rather than emulating their work, we chose to use every

hospitalization instance which resulted in more data for training; we wanted to try Deep Learning methods which also requires large amounts of data. Moreover, clinical utility in the real world is more accurately represented by using all hospitalizations which a patient may have as health providers would be interested in predicting risk for any instance rather than a single or random hospitalization, as others have done.

Our outcome was unplanned, all-cause readmission within 30 days of a reference hospitalization. 30-day windows are considered directly related to the initial hospitalization and are a key health policy metric in evaluating and improving healthcare quality[5,99,115,116]. Furthermore, 20% (median 12 days) of patients with HF are readmitted within 30 days of a discharge in which only a third of readmissions were primarily due to HF[117]; indeed most readmissions are for non-HF reasons[117,118]. Each reference hospitalization occurring after the first echocardiogram was used to predict a subsequent 30-day all-cause readmission; a readmission could become a reference hospitalization if there was a subsequent admission within 30 days. Others[42,113] used only HF related hospitalizations as their reference hospitalization with unremarkable ML prediction performance; instead, we used a previously defined cohort of patients with HF and considered all hospitalizations as potential reference hospitalizations since patients with HF are frequently hospitalized for non-HF related issues as mentioned above.

Machine Learning Methods

We used common ML algorithms and approaches[11,14,17,42,50] to train our models. Our outcome-labelled data was split into development (80%) and validation (20%) sets by patients and hospitalizations such that no patients in the development set were included in the validation set. The development set was used to train and test various ML algorithms (XGBoost, Gradient Boosting Machine [GBM] , AdaBoost, CatBoost, Light GBM, Linear Support Vector Classifier [SVC], Gaussian Naïve Bayes, Random Forest, Decision Tree, L1 logistic regression, and Neural Network) and tune model hyperparameters using k-fold (k=10) cross validation[39]. The modeling process was started with the default set of hyperparameters provided by SkLearn[119], which has shown good predictive performance on various data sets; for XGBoost[120] we used max_depth (maximum tree depth for base learners; range 3-10),

n_estimators (number of gradient boosted trees; range 10-100), scale_pos_weight (balancing of positive or negative weights; range none or calculated as follows [(total number of observations - number of positive observations)/number of positive observations]) and other default parameters. As well, we used PyCaret[121], Auto-keras[122] and H20 Driverless[123], all of which automatically optimize a range of hyperparameters.

To address the correlations between multiple hospitalizations within the same patient, which some consider a limitation[37], we modeled hospitalization history of patients as a time ordered series with a LSTM[124] (long short-term memory recurrent neural network) using the same architecture as others did for predicting readmissions in the HF population[125]; our data set was first transformed into a longitudinal set, then LSTM was applied.

We used the validation set to evaluate prediction performance of the ML models and LaCE score

Predictors

Predictor variables included those that were informed by the literature[4,42,126] and those that incorporated information from our datasets (Table 4.3 and Figure 4.5). We included feature groupings based on demographic information, hospital admission characteristics, healthcare utilization (number of visits to hospitals and physicians), drug utilization (measured at the time of reference hospitalization and 30 days prior), co-morbidity history[87,127], history of cardiac procedures, and laboratory test results (most recent one completed prior or during hospitalization). Health care utilization features were binned into categories. As well, we incorporated time sensitive variables for each hospital discharge such as time elapsed since last cardiac related procedure, number of hospitalizations, physician and emergency department visits in the previous 6 months.

Missing Data and Outliers

We anticipated missing data when handling laboratory results. It should be noted that the Provincial Laboratory fully captures all performed laboratory tests and that any “missing” data are laboratory tests that were not ordered. Missing laboratory data is classified as either missing at random or missing not at random[45,50] and imputation methods are not favourable

with ML prediction tools intended for deployment in a work setting[52]. Instead, we included missing indicator variables as others have done[46,51,53] although we understand this use is controversial[128]. However, if the models are to be deployed in real-world settings, the missing indicator approach is most practical. Tree-based ML algorithms (e.g. XGBoost) are able to handle missing data better than regression based algorithms[50]. We made use of the *Sparsity-aware split finding* feature[95] when training the XGBoost algorithm to address missing data without the need for missing indicators. All other ML algorithms required the use of missing indicators. Other HF ML studies excluded missing medical data or used imputation to handle missing data, both recognized as leading to high bias[42,107,113].

We encountered outliers when analyzing healthcare utilization features such as number of previous physicians, hospital or emergency department visits. In these cases, we identified outliers if they were more than 1.5xIQR (interquartile range) above the 3rd quartile and binned them into the highest category.

Analysis and Prediction Evaluation

We first performed a descriptive analysis comparing those with and without a readmission and those in the training and validation sets using chi-square tests and t-tests. **This descriptive analysis was done at the patient level in which patients could experience multiple hospitalizations each.** We also included number of events, final sample size, and distribution of missingness in the lab data.

Using the validation set, we reported prediction performance metrics[11,17,19,20,22] of the ML model with the highest AUROC (area under the receiver operating characteristics) curve and compared them with those of the LaCE score. This included positive likelihood ratios (LR+)[20], true/false positives, true/false negatives, category-less net reclassification improvement (NRI)[111,129] and positive predictive value (PPV, equivalent to post-test probability); these were stratified by prediction thresholds of the models with the exception of NRI. For binary classification studies like ours, AUROC curves correspond to c-statistics which are a measure of model discrimination performance, the extent to which a model predicts a higher probability of an outcome among patients who actually had the outcome compared to those who did not[19].

LR+'s are used as a multiplier to convert pre-test odds (or probabilities) to post-test odds (or probabilities) thus providing some measure of clinical informativeness of the ML model. The NRI measures the amount of correct reclassification[111] (predicted risk moving upward for events and downward for non-events) when our defined outcome for the ML model was compared to that of the LaCE. Because there are no established risk categories for 30-day readmissions, we used category-less NRI[129] for comparing ML prediction to LaCE; an NRI above zero indicates better risk prediction for the ML model compared to the LaCE score.

We also included calibration plots[19] for the ML model; calibration is perhaps considered the most important property of a model and reflects the extent to which predicted values align with observed values and is most often illustrated by a plot of observed vs predicted[18,19]. Along with the calibration plot, we added a negative predictive value (NPV) vs. predicted risk plot to highlight the relationship between low predicted risk and true negatives (those who did not experience the outcome).

Because ML models do not estimate an interpretable quantity relating predictors to outcomes (not the purpose of ML prediction), it is not appropriate to summarize that relationship with a single parameter; instead, the influence, or impact, of individual predictors can be summarized using variable importance which is a rank-ordering of variables which are most important for the ML model's prediction performance[50]; variable importance does not have a causal or statistical meaning. To address interpretability[59] of our ML model, we reported feature importance[50] and feature impact using SHAP plots[91,92] for the ML model with the highest discrimination performance using the validation set. As well, we reported AUROCs for different combinations of features to see the effect of reduced predictors on discrimination performance using the ML model with the highest overall AUROC.

In sub-group analyses, we assessed the discrimination performance of the ML model according to type of HF in the full dataset: heart failure with reduced ejection fraction (HFrEF), heart failure with mid-range ejection fraction (HFmEF) and heart failure with preserved ejection fraction (HFpEF). As well, we performed an analysis using only HF specific reference hospitalizations identified by the main diagnosis field (using ICD-10 code I50) and estimated the

discrimination performance (c-statistic); we also reported the distribution of missing laboratory test values in this subset.

All findings were reported using TRIPOD[68] and other guidelines[11,17] specific to ML projects. All analyses were done using Python (version 3.6.8, Python Software Foundation), SciKit Learn[94] (version 0.23.2), SHAP[92] (version 0.35), XGBoost[95] (version 0.90), Pandas[96] (version 1.0.5) and STATA/MP V.15.1 (StataCorp). This study received ethics approval from the University of Alberta ethics board (Pro00097809).

Results

We identified 9,845 patients with HF representing 48,745 reference hospitalizations for our study period (Figure 4.1). As mentioned in the Methods section, our unit of analysis is at the hospitalization level, thus, each patient could have multiple admissions and be represented in multiple instances. Most of the patients in our dataset had 7 or fewer admissions (Figure 4.6). The top 2 most frequent primary diagnoses for both reference admissions and readmissions were heart failure and COPD exacerbations (Table 4.4). As expected, there were differences between those who were readmitted and those who were not (Table 4.5) while the distribution of characteristics was similar between those in the development and validation sets (Table 4.6). The mean age at reference admission was 71.5 (SD=14); males accounted for 5,539 (56%) of HF patients. Those with hypertension, prior ischemic heart disease, renal disease and depression represented 96.2% (n=9,471), 90.3% (n=8,894), 52.0% (n=5,112), and 61.7% (n=6,078) of patients with HF, respectively.

10,182 (20.9%) reference hospitalizations were followed by an unplanned 30-day readmission after discharge. Our development and validation sets had 7,876 (80%) and 1,969 (20%) patients corresponding to 39,066 (80%) and 9,679 (20%) reference hospitalizations, respectively. There were 8,101 (20.8%) and 2,081 (21.1%) readmissions in the development and validation sets, respectively.

Missing laboratory data in our dataset ranged from 22-83% (Table 4.7). The laboratory measurements with the highest missing values were BNP (83.1%) and NT-proBNP (74.0%),

which are not commonly ordered in practice, particularly in the earlier years of the cohort in Alberta, Canada.

With respect to our ML models, from our validation set, the boosted tree-based ML algorithms had the highest AUROCs with XGBoost being the highest at 0.65 while the LaCE score was at 0.57 (Table 4.1, Figure 4.2). Also, the LSTM classifier used to model hospitalizations at the patient level did not perform as well as the boosted trees (Table 4.1). Calibration plots for the XGBoost classifier showed that predicted risk of readmission was aligned with observed risks and that low predicted risks were associated with fewer actual outcomes highlighting higher negative predicted values at lower predicted risks (Figure 4.3). Above a predicted risk of around 0.55, what few hospitalizations were present all led to an actual readmission, thus the calibration in this segment illustrates that the ML classifier underestimated the predicted risk.

When we stratified across predictive thresholds for the XGBoost classifier from the low to high end of predicted probabilities, the LR+ ranged from 1.00-6.12 and the PPV (post-test probability) from 0.21-0.62 (Table 4.2). Similar stratification of the LaCE score resulted in LR+ values ranging from 1.00-1.20 and PPVs from 0.21-0.24 (Table 4.8). These can be contrasted with the pre-test probability (using prevalence) of 20.9% observed in our data. Further comparison yielded a NRI of 0.34 (95% CI 0.30-0.40) which indicates that a higher proportion of patients were correctly reclassified with the XGBoost classifier compared to the LaCE score.

Regarding interpretability and feature importance of the XGBoost classifier, previous hospitalization history and hemoglobin levels carried the highest impact for predicting readmissions (Figure 4.4) with SHAP plots indicating that higher number of previous hospitalizations and lower hemoglobin levels predicted higher risk of readmissions (Figure 4.7).

Reducing the number of predictors also reduced the discrimination performance in our XGBoost model. Using only laboratory test results, admission characteristics, or drug utilization (Table 4.3) resulted in AUROCs of 0.595, 0.574, and 0.558 respectively (Table 4.9).

In the sub-group analysis, most patients (n=5,702; 57.92%) were classified as HFpEF and estimated c-statistics were 0.67, 0.61 and 0.63 across HFrEF, HFmEF and HFpEF, respectively

(Table 4.10). With respect to ML prediction in the subset of HF specific reference hospitalizations, the estimated c-statistic was 0.60 (Table 4.10) and the distribution of missing lab values was like that of the entire data set (Table 4.11).

Discussion

In this study, we used administrative health data to develop and validate ML models which predicted the risk of unplanned readmissions 30 days after discharge and compared prediction metrics with the widely used LaCE score. The ML approach can leverage larger amounts of data and include more predictors offering an advantage over non-ML methods like the LaCE score, which is a score card consisting of 4 predictors (length of stay, age, Charlson score and number of emergency department visits); indeed, ML models can go beyond a simple score card by incorporating many predictors. The LaCE score provided minimal predictive capabilities and was not more informative than the pre-test probability of 20.9%. Although the ML models we tested performed better than the LaCE score, even our best performing ML model (the XGBoost classifier) was not sufficiently informative as a classifier to predict risk of readmissions with LR+ ranging from 1.00 to 6.12. A LR+ >10 is considered strongly informative with conclusive changes from pre-test to post-test probabilities [20]. However, because our ML classifier was more informative than the LaCE score, which is currently being used in many health-systems including in Alberta, it is possible that health-system administrators may find value with our approach of including feature importance and feature impact (SHAP plots) to permit a measure of interpretability (in terms of which variables are most strongly associated with readmission risk) not commonly seen in other ML studies. However, it should be noted that interpretability is not usually a consideration of ML prediction.

Our results indicate that predicting readmissions in patients with HF is a difficult undertaking, whether using ML or non-ML methods. Indeed, whether we considered all-cause or HF specific reference hospitalizations or HF status, ML prediction was unremarkable. The varying and somewhat uninformative predictive performance across ML models reflects this difficulty which may be explained by data quality as it is estimated that over 80% of the work done in ML projects is comprised of data preparation[32]; there was not enough variation in the administrative data for substantial improvements in ML prediction. To augment ML prediction,

administrative health data could be linked to social factors data, which are known determinants of health, however, this cannot be done in many jurisdictions including Alberta. The purpose of this study was to see if regularly captured administrative data could train ML models to predict readmissions. ML models trained on administrative data alone cannot be expected to outperform non-ML methods that also use administrative data alone, especially when many of the features are generated using clinically guided expertise[16,34]. Incorporating the entire taxonomy of "big health data" would likely improve ML prediction by linking biological, geospatial, electronic health records, personal monitoring device, and effluent sources of data[34] as all of these contribute to patient heterogeneity. Furthermore, it is well known that the elderly are a heterogeneous segment of society with regards to health[130,131]. The HF population also inherently shares this heterogeneity in which administrative health data alone cannot explain the variation in readmission risk. Indeed, patients with HF are already at high risk of readmissions and further ML risk stratification in this already high-risk group was not possible using our datasets; there are other unidentified, or rather, uncaptured factors which will influence prediction performance.

Other ML studies[42,105,114,132] which looked at readmissions in patients with HF had AUROCS ranging from 0.61-0.67 with rates of readmissions around 21%, similar to our findings. However, our study also included recommended metrics like LR+, pre-test/post-test probabilities across prediction thresholds and NRI to compare classifiers showcasing that our ML classifier was not substantially informative. We also trained newer ML algorithms (e.g., XGBoost) using Alberta specific administrative data and provided an assessment of ML model interpretability using feature importance and impact plots. Our finding that ML classifiers better predict readmissions when compared to variations of the LaCE score (or more conventional regression-based models) is also noted in ML studies by Frizzell et al.[42] , Bayati et al[132] and Shin et al[107]. Overall, our study's findings are aligned with others in that ML prediction of readmissions in patients with HF does not carry much clinical utility[42].

Our study benefited from complete records of hospitalizations, emergency visits, claims, and medication history anywhere in the province of Alberta (not just within the site of the reference hospitalization) using administrative data validated by Canadian Institute for Health

Information[110]. Although our data cannot capture events outside of Alberta, we expect this scenario to represent very few instances which will not affect ML training. However, we did have substantial missingness with laboratory data, which may influence prediction performance. Even when we analyzed the subset of HF specific reference hospitalizations, there was a similar distribution of missingness. It should be noted that these were not necessarily missing, but simply not completed on a patient. As a result, these values were not available to health providers either during the clinical decision-making process. Others[42] simply excluded missing data from their ML training data and identified this method as leading to high bias[107] while we decided to use XGBoost's capabilities in handling missing values. Furthermore, our study required 2 echocardiograms to verify HF status, and may not be representative of all HF patients. ML projects are entirely data driven and even though we incorporated around 160 different predictors using administrative data, we were not able to measure many predictors which substantially contribute to readmissions, namely, social factors. Frizzell et al. also noted this point and went further to highlight that even with social factors data that ML prediction performance is not guaranteed[42]. How predictive these social factors are remains uncertain, but some improvement in performance could be expected with their inclusion[118]. This was evident as our dataset did not sufficiently predict readmissions in an informative manner. Similar to other ML prediction studies[37], we assumed all hospitalizations were independent events and thus we ignored correlations within patients hospitalized multiple times, which may have unduly influenced the analysis (although we had a large sample). Nevertheless, correlation in datasets like ours leading to poor ML prediction has not been substantiated in the literature. Another study[42] looking at the same outcome simply took the first hospitalization for patients with multiple hospitalizations and did not have sufficient prediction performance. We trained an LSTM model to address this, however, it had lower discrimination performance compared to the findings of Ashfaq et al[125]. As with all ML projects, our models were trained using local data and may not be generalizable however, the data was derived from an overall population of 4 million patients in Alberta and should be broadly representative of most HF patients in other settings. Moreover, a benefit of the ML process is that models can be easily retrained to other populations.

This study suggests that ML methods may better predict readmissions compared to non-ML methods like the LaCE score, although still at an insufficient level to be of value to many health system planners. Health systems should consider our results, and those of others, if they wish to deploy ML technology into their workflows. While our ML methods had limited clinical prediction ability when trained only on administrative data, ML is just one technique of Artificial Intelligence (AI), and other techniques like Natural Language Processing (NLP) may add more prognostic information by processing practitioner chart notes. Indeed, in building an effective data driven model, all data available in health institutions (e.g., patient chart data, administrative data) should be considered to improve predictive capabilities. Combined, ML and NLP may provide a model with high enough prediction performance for health practitioners to use in clinical settings. Deployment of ML models in general requires technology readiness[133], IT infrastructure, and especially evaluations of ML impact on outcomes, which is severely lacking[56]. While the ML process may be promising, health systems must acknowledge barriers to data as many jurisdictions do not permit widespread use of administrative health data and data sharing is a major issue in almost all health systems. Indeed, having access to data detailing outcomes and patient histories, as well as other types of personal non-administrative sources of data may improve ML prediction performance and utility and should be explored."

Table 4.1. Discrimination performance of ML models and LaCE score using AUROCs.

Classifier	Development set	Validation set
XGBoost Classifier	0.685	0.654
Gradient Boosting Machine (GBM) Classifier	0.687	0.650
AdaBoost Classifier	0.655	0.646
CatBoost	0.851	0.642
Light GBM Classifier	0.811	0.641
Linear SVC	0.671	0.639
Gaussian Naïve Bayes	0.638	0.624
Random Forest Classifier	1.000	0.617
Decision Tree Classifier	0.741	0.597
Logistic Regression (L1)	0.591	0.596
Neural Network Classifier	0.579	0.578
LSTM	0.681	0.624
LaCE	0.570	0.570

AUROC: Area under the receiver operating characteristic curve; SVC: Support Vector Classifier; LSTM: Long short-term memory recurrent neural network

Table 4.2. Prediction metrics for the XGBoost classifier across predictive thresholds.

Predictive Threshold	TP	FN	FP	TN	PPV (post-test probability*)	LR+
0	2081	0	7782	0	0.21	1.00
0.1	2065	16	7464	318	0.22	1.03
0.2	1371	710	3456	4326	0.28	1.48
0.3	477	1604	797	6985	0.37	2.24
0.4	95	1986	126	7656	0.43	2.82
0.5	18	2063	11	7771	0.62	6.12
0.6	0	2081	0	7782	(--)	(--)

*Compared to a pre-test probability of 20.9% using prevalence

TP: true positives; FN: false negatives; FP: false positives; TN: true negatives

PPV: positive predictive value; LR+: positive likelihood ratio (this is a multiplier used to convert a pre-test odds to post-test odds and subsequently, pre-test probability to post-test probability)

(--): no value

Figure 4.1. Patient flow diagram of study participants.

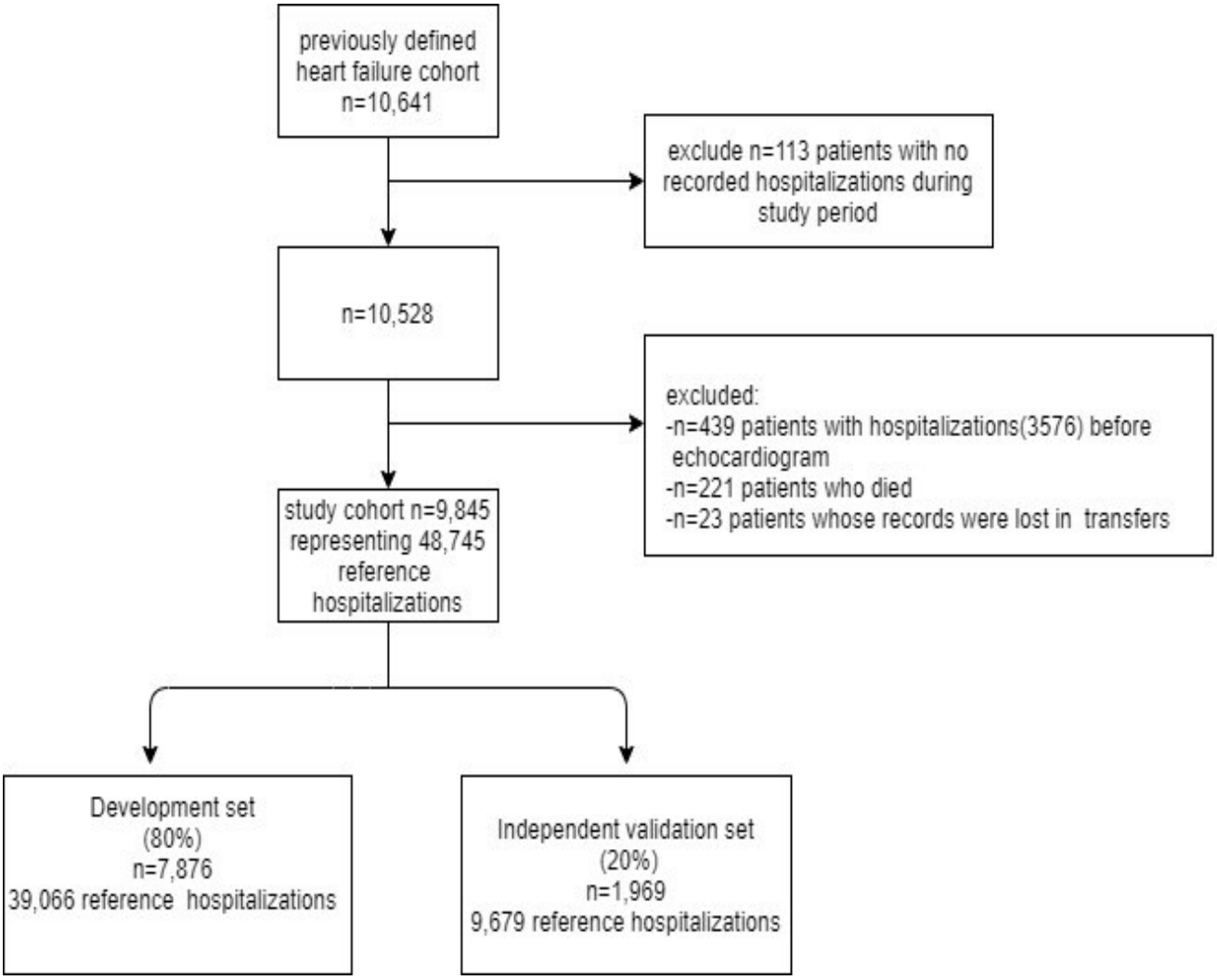
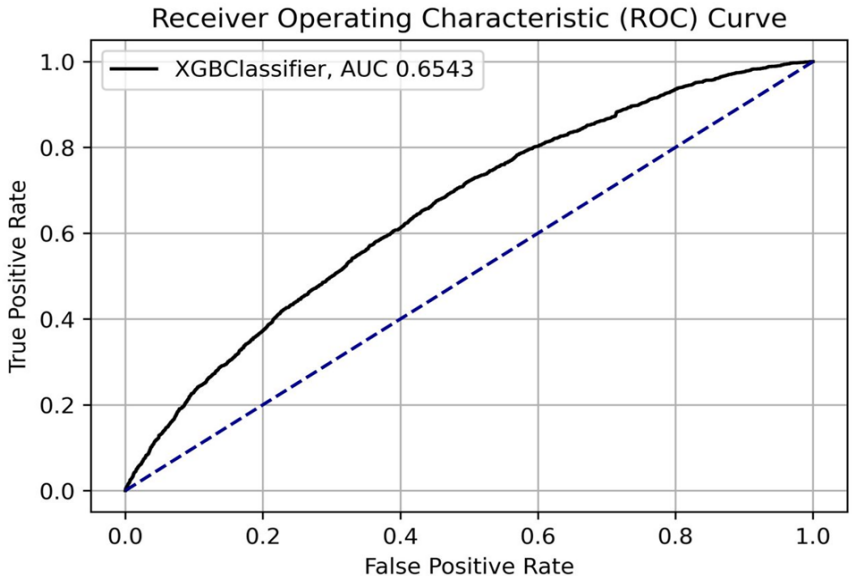


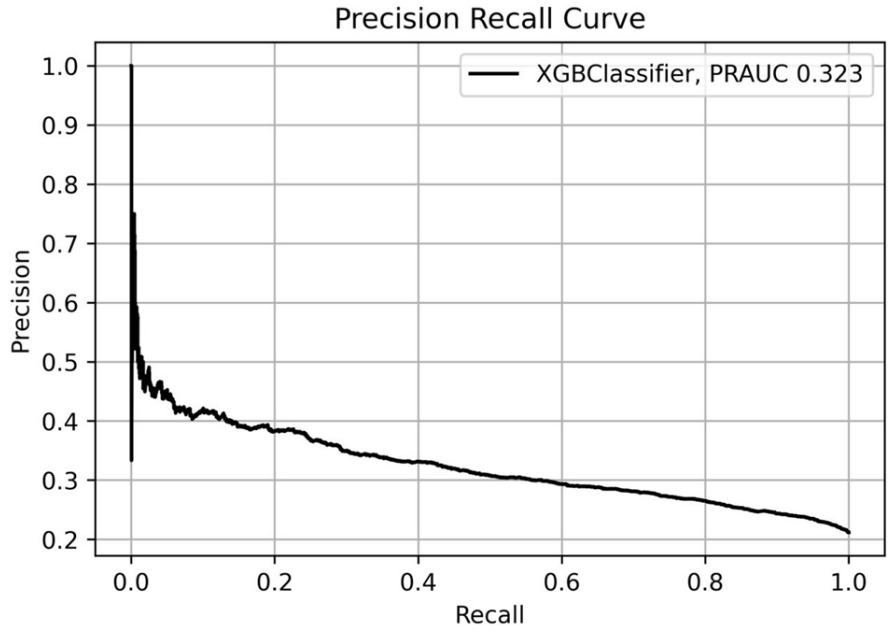
Figure 4.2. Area under receiver operating characteristic (A) and precision-recall curve (B) for the XGBoost model.

(A)



XGBClassifier: XGBoost Classifier; AUC: area under the curve

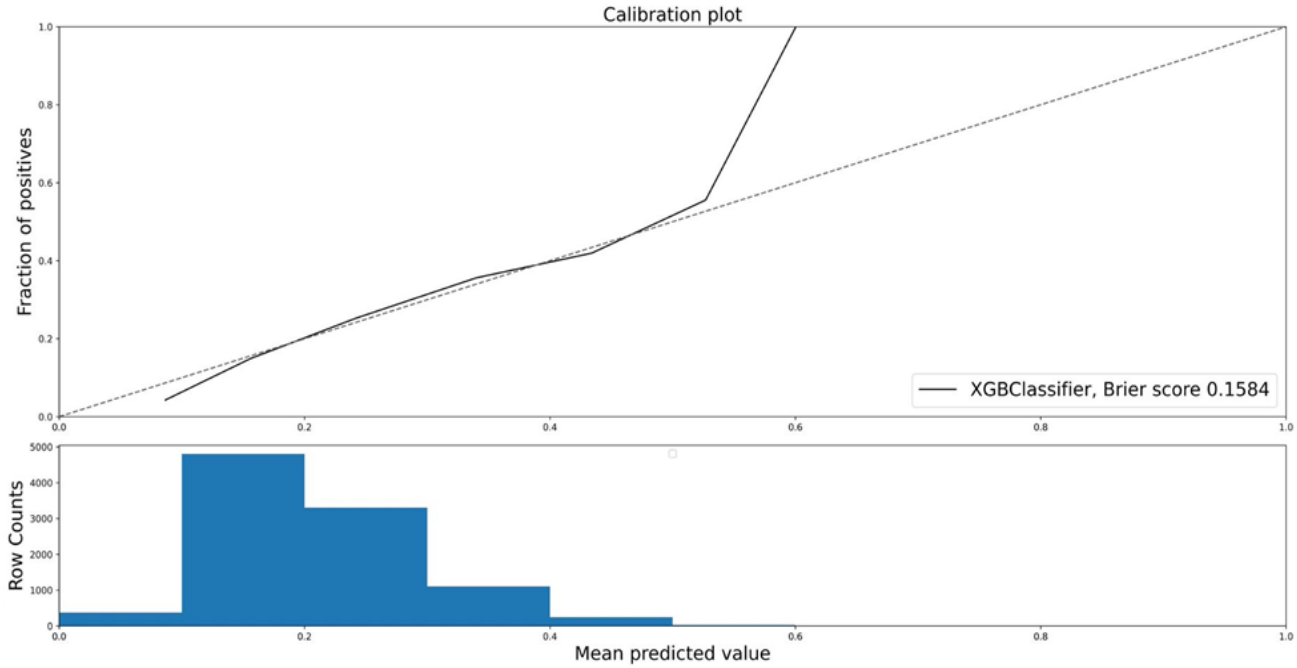
(B)



PRAUC: area under the precision-recall curve

Figure 4.3. Calibration plot with counts of hospitalizations (A) and NPV (negative predictive value) vs predicted risk (B) for the XGBoost classifier.

(A)



(B)

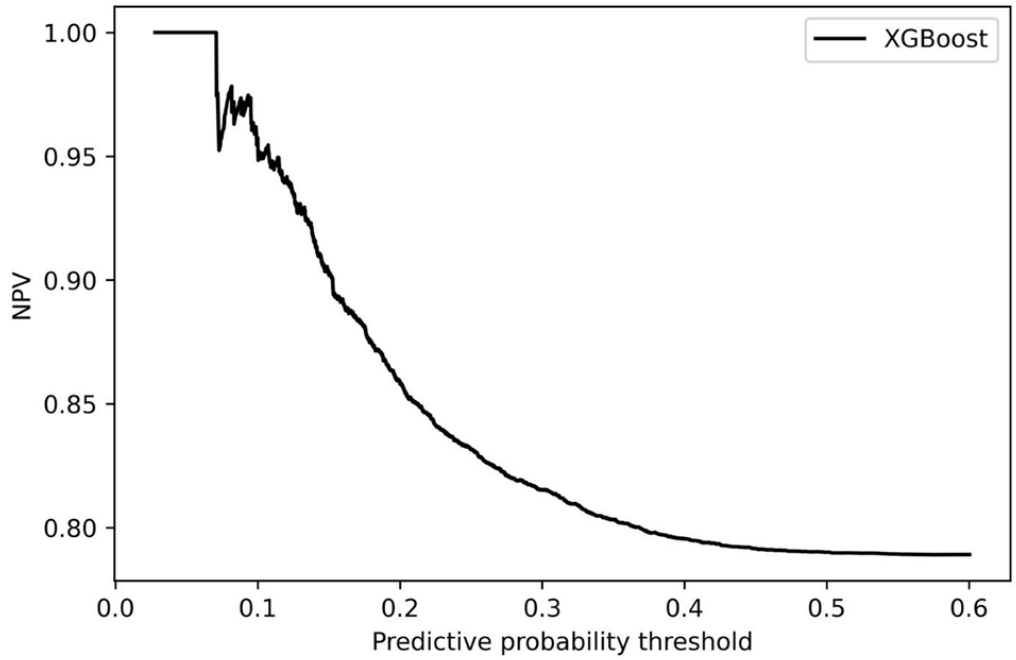
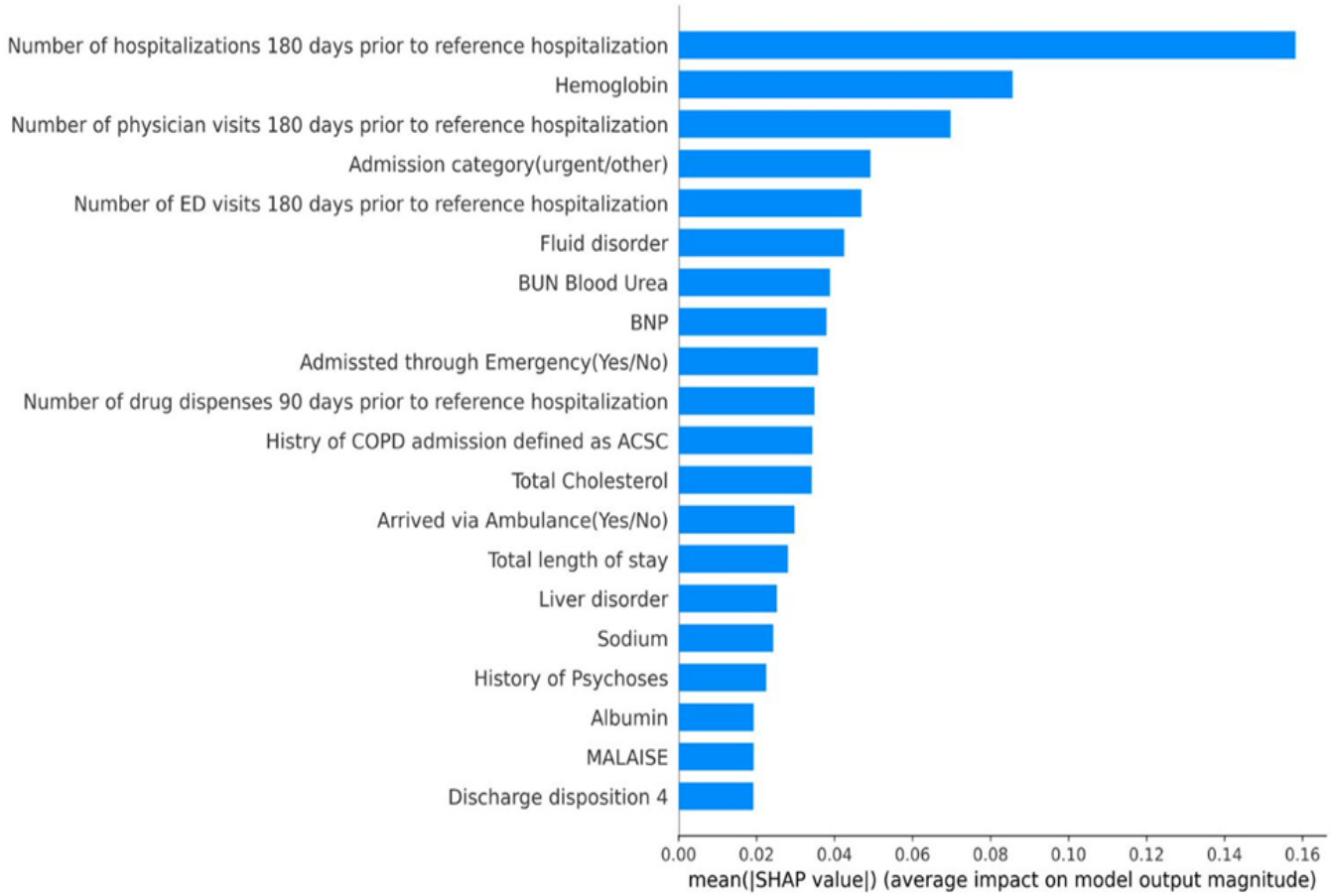


Figure 4.4. Feature importance for the XGBoost classifier.



Note: the x-axis represents variable importance and impact on ML model output which is specific to the individual algorithm (in this case, XGBoost) and does not have a statistical interpretation. Variable importance is a relative comparison of feature contribution to the overall prediction of the outcome (predictors are rank-ordered); predictors are arranged from top to bottom in order of decreasing impact on ML model output; BNP: B-type natriuretic peptide; ACSC: ambulatory care sensitive condition; COPD: chronic obstructive pulmonary disease; Discharge disposition 4: discharged to private home with community supports

Appendix to Chapter 4

Table 4.3. Predictors used for model training (data dictionary).

Category	Predictor
Demographics	Age Sex Income Rural/urban Forward sortation index (postal codes)
Admission characteristics	Main reason for admission (ICD-10) Resource intensity weight (allocation of resources for reference hospitalization) length of stay Admission category (urgent or other) Admitted through emergency (yes/no) Admitted through ambulance (yes/no) Co-morbidity that occurred during hospitalization (yes/no) Facility ID Institution to Institution from Discharge disposition Calendar year of admission
Healthcare Utilization	Number of physician visits in previous 90 days** Number of pharmacy visits in previous 90 days Number of prescription dispenses in previous 90 days Number of hospitalizations in previous 180 days Number of physician visits in previous 180 days** Number of emergency visits in previous 180 days Number of general practitioner visits in previous 180 days** Number of cardiologist visits in previous 180 days**
Ambulatory sensitive care conditions	 Heart failure and pulmonary edema Angina COPD Asthma Hypertension Epilepsy
Comorbidity group*	Alcoholism Arrhythmia Anemia Heart failure and pulmonary edema Cancer

	Lymphoma Coagulopathy Iron deficiency anemia Depression Diabetes Substance abuse Fluid and electrolyte disorder HIV infection Hypertension Hypothyroidism Liver disorder Neurologic disorder Obesity Paralysis Psychoses Peptic ulcer disease Pulmonary circulation disorder Pulmonary disease Peripheral vascular disease Renal disease Rheumatoid arthritis Valvular disease Weight loss Stroke Prior ischemic heart disease Dyslipidemia Tobacco Cognitive impairment Delirium Dementia Difficulty walking Falls Incontinence Gait abnormalities Senility Vascular dementia Pressure ulcer Malaise
Cardiac procedure group	 echocardiogram ejection fraction Date of echocardiogram Implantable cardioverter defibrillator Cardiac resynchronization therapy

	Left ventricular assist device Percutaneous coronary intervention Coronary artery bypass Aortic valve procedure Mitral valve procedure Procedure date
Drug Utilization (at the time of reference hospitalization and 30 days prior)	Level 3 or 4 Anatomical Therapeutic Chemical code Ace inhibitors Beta blockers Mineral corticosteroids Sodium-glucose cotransporter-2 (SGLT2) hydralazine Isosorbide Dinitrate digoxin calcium channel blockers anti-arrhythmic agents dyslipidemia drugs diuretics
Laboratory tests	Blood urea nitrogen MCV Hemoglobin Potassium Creatinine eGFR Sodium Hematocrit ALT Ferritin Serum albumin A1C BNP NT/proBNP Number of days between lab test and reference hospitalization

*Determined using ICD-10 (International Statistical Classification of Diseases and Related Health Problems); co-morbidities were derived from all Claims data and hospitalizations prior to and including the reference hospitalization.

**These features were derived from Physician Claims data

Table 4.4. Admission statistics and main reason for admission.

Admission statistics	Number (% of total reference hospitalizations)
Number of reference admissions	48745 (100)
Number of 30-day readmissions	10182 (20.9)
Top 5 most frequent main diagnoses for reference hospitalizations (ICD-10):	
Heart failure	7574 (15.5)
COPD exacerbation	1703 (3.5)
COPD with lower respiratory tract infection	1545 (3.2)
Physical therapy (breathing exercises)	1298 (2.7)
Pneumonia	1148 (2.3)
Top 5 most frequent main diagnoses for readmissions (ICD-10)	
Heart failure	1901 (18.7)
COPD exacerbation	349 (3.5)
COPD with lower respiratory tract infection	325 (3.2)
Pneumonia	231 (2.3)
Acute renal failure	193 (1.9)

ICD: International Statistical Classification of Diseases and Related Health Problems
COPD: Chronic obstructive pulmonary disease

Table 4.5. Characteristics of those who were readmitted and those who were not. This descriptive analysis was done at the patient level in which patients could experience multiple hospitalizations each.

Characteristic	Patients without Readmission (n=9,749)* Number~	%	Patients with Readmission (n=4,663)* Number~	%	p-value
Age at reference admission:					
Mean (SD)	71.57 (14.0)	**(--)**	71.32 (14.0)	**(--)**	**(--)**
18-45	492	5.0	174	3.7	<0.001
45-65	2544	26.1	1025	22.0	<0.001
>65	7402	75.9	3568	76.5	0.44
Sex:					
Male	5492	56.3	2544	54.6	0.04
Female	4257	43.7	2119	45.4	0.04
Discharge disposition:**					
2	1405	14.4	425	9.1	<0.001
4	5084	52.1	2059	44.2	<0.001
5	8324	85.4	3267	70.1	<0.001
6	360	3.7	181	3.9	0.58
12	38	0.4	8	0.2	0.03
30	182	1.9	37	0.8	<0.001
40	181	1.9	53	1.1	0.00
62	31	0.3	8	0.2	0.11
Medical History***					
Ejection fraction, mean (SD)	48.46 (13.07)	(--)	48.83 (13.05)	(--)	(--)
Hypertension	9367	96.1	4517	96.9	0.02
Arrhythmia	9214	94.5	4393	94.2	0.46
Ischemic heart disease	8783	90.1	4217	90.4	0.52
Depression	5971	61.2	2978	63.9	0.00
Injury	9293	95.3	4474	95.9	0.09
Dyslipidemia	6794	69.7	3255	69.8	0.89
Rheumatoid arthritis	5904	60.6	2901	62.2	0.06
Alcohol use disorder	3079	31.6	1569	33.6	0.01
Neurologic disorder	8739	89.6	4220	90.5	0.11
Coagulopathy	3193	32.8	1568	33.6	0.30
Congestive heart failure	9578	98.2	4452	95.5	<0.001

Pulmonary disease	6804	69.8	3388	72.7	<0.001
Poisoning	2103	21.6	1075	23.1	0.04
Pulmonary circulation disorder	4045	41.5	2065	44.3	0.00
Deficiency anemia	4525	46.4	2361	50.6	<0.001
Liver disorder	1588	16.3	856	18.4	0.00
Peptic ulcer disease	1449	14.9	798	17.1	<0.001
Renal disease	5006	51.3	2616	56.1	<0.001
Blood loss anemia	720	7.4	430	9.2	<0.001
Valvular disease	4653	47.7	2176	46.7	0.23
Fluid and electrolyte disorder	7215	74.0	3661	78.5	<0.001
Diabetes	5666	58.1	2860	61.3	<0.001
Stroke	2822	28.9	1288	27.6	0.10
Obesity	3192	32.7	1569	33.6	0.28
Cancer	3706	38.0	1804	38.7	0.44
Peripheral vascular disease	4329	44.4	2172	46.6	0.01
Weight loss	2566	26.3	1301	27.9	0.05
Hypothyroidism	2988	30.6	1487	31.9	0.13
Drug abuse disorder	907	9.3	497	10.7	0.01
Paralysis	750	7.7	344	7.4	0.50
Lymphoma	698	7.2	365	7.8	0.15
Smoking	356	3.7	194	4.2	0.14
HIV	41	0.4	23	0.5	0.54
Psychoses	2313	23.7	1227	26.3	<0.001
Mean length of stay of reference hospitalization (SD)	16 (31.6)	**(--)**	15.8 (28.2)	**(--)**	0.58
Mean LaCE score (SD)	59.9 (6.5)	**(--)**	61.4 (5.3)	**(--)**	<0.001

~Unless otherwise specified

*Patients can be in either category because those with multiple hospitalizations may have experienced a readmission from some reference hospitalization and no readmission from others.

**2: transferred to continuing care; 4: discharge to private home with supports from community; 5: discharge to private home without supports; 6: sign out, absent without leave; 12: patient did not return from pass; 30: transfer to long-term care home, mental health and/or addiction centre or hospice; 40: transferred to assisted living; 62: left against medical advice

***derived using ICD-10 codes (International Statistical Classification of Diseases and Related Health Problems)

Table 4.6. Characteristics of patients in development and validation sets. This descriptive analysis was done at the patient level in which patients could experience multiple hospitalizations each.

Characteristics	Development set (n=7,876) Number	%	Independent validation set (n=1,969) Number	%	p value
Age at reference admission:					
Mean (SD)	71.5 (14.0)	**(--)**	71.5 (13.9)	**(--)**	**(--)**
18-45	387	4.9	107	5.4	0.34
45-65	2062	26.2	497	25.2	0.40
>65	6008	76.3	1497	76.0	0.81
Sex:					
Male	4419	56.1	1120	56.9	0.54
Female	3457	43.9	849	43.1	0.54
Discharge disposition*:				0.0	
2	1216	15.4	311	15.8	0.70
4	4430	56.2	1104	56.1	0.89
5	6866	87.2	1724	87.6	0.65
6	403	5.1	88	4.5	0.24
12	39	0.5	5	0.3	0.15
30	172	2.2	27	1.4	0.02
40	159	2.0	36	1.8	0.59
62	34	0.4	1	0.1	0.01
Medical History:**					
Ejection fraction, mean (SD)	48.60 (13.10)	(--)	48.3 (13.0)	(--)	(--)
Hypertension	7575	96.2	1896	96.3	0.81
Arrhythmia	7462	94.7	1854	94.2	0.30
Prior ischemic heart disease	7115	90.3	1779	90.4	0.99
Depression	4849	61.6	1229	62.4	0.49
Injury	7508	95.3	1887	95.8	0.33
Dyslipidemia	5471	69.5	1398	71.0	0.18
Rheumatoid arthritis	4784	60.7	1198	60.8	0.93
Alcohol use disorder	2518	32.0	616	31.3	0.56
Neurologic disorder	7084	89.9	1760	89.4	0.46
Coagulopathy	2617	33.2	659	33.5	0.84
Congestive heart failure	7753	98.4	1946	98.8	0.20
Pulmonary disease	5542	70.4	1354	68.8	0.17
Poison	1694	21.5	445	22.6	0.29

Pulmonary circulation disorder	3294	41.8	855	43.4	0.20
Deficiency anemia	3695	46.9	920	46.7	0.88
Liver disorder	1315	16.7	315	16.0	0.46
Peptic ulcer disease	1165	14.8	309	15.7	0.32
Renal disease	4070	51.7	1042	52.9	0.32
Blood loss anemia	590	7.5	157	8.0	0.47
Valvular disorder	3756	47.7	976	49.6	0.14
Fluid and electrolyte disorder	5849	74.3	1490	75.7	0.20
Diabetes	4580	58.2	1155	58.7	0.68
Stroke	2306	29.3	554	28.1	0.32
Obesity	2584	32.8	639	32.5	0.76
Cancer	3050	38.7	749	38.0	0.58
Peripheral vascular disease	3495	44.4	890	45.2	0.51
Weight loss	2112	26.8	554	28.1	0.24
Hypothyroidism	2443	31.0	597	30.3	0.55
Drug abuse disorder	743	9.4	179	9.1	0.64
Paralysis	630	8.0	130	6.6	0.04
Lymphoma	561	7.1	155	7.9	0.25
Smoking	298	3.8	64	3.3	0.26
HIV	30	0.4	11	0.6	0.27
Psychoses	1936	24.6	448	22.8	0.09
Mean length of stay of reference admission (SD)	16.0 (30.6)	**(--)**	15.9 (31.9)	**(--)**	0.89
Mean LaCE score (SD)	60.1 (6.3)	**(--)**	60.3 (6.4)	**(--)**	0.21
Mean number of multiple hospitalizations (SD)	4.94 (4.0)	**(--)**	5.00 (4.2)	**(--)**	0.55

*2: transferred to continuing care; 4: discharge to private home with supports from community; 5: discharge to private home without supports; 6: sign out, absent without leave; 12: patient did not return from pass; 30: transfer to long-term care home, mental health and/or addiction centre or hospice; 40: transferred to assisted living; 62: left against medical advice

**derived using ICD-10 codes (International Statistical Classification of Diseases and Related Health Problems)

Table 4.7. Distribution of missing laboratory data.

Lab Test	Entire Dataset*		Development Set**		Validation Set***		p-value^
	% Missing	Number missing	% Missing	Number Missing	% Missing	Number Missing	
Blood Urea Nitrogen	25.5	12,428	25.4	9,935	25.8	2,493	0.51
MCV	22.0	10,748	22.0	8,594	22.3	2,154	0.59
Hemoglobin	22.0	10,745	22.0	8,592	22.2	2,153	0.60
Total Cholesterol	52.2	25,423	52.0	20,326	52.7	5,097	0.27
Potassium	22.1	10,764	22.0	8,604	22.3	2,160	0.53
Creatinine	22.1	10,761	22.0	8,602	22.3	2,159	0.54
eGFR	22.4	10,942	22.4	8,749	22.7	2,193	0.58
Cholesterol Ratio	60.8	29,625	60.5	23,624	62.0	6,001	0.01
Sodium	22.1	10,765	22.0	8,605	22.3	2,160	0.54
Hematocrit	22.0	10,747	22.0	8,593	22.3	2,154	0.58
ALT	27.7	13,486	27.6	10,772	28.0	2,714	0.36
Ferritin	50.6	24,671	50.5	19,733	51.0	4,938	0.37
Albumin	36.8	17,918	36.8	14,370	36.7	3,548	0.82
A1C	46.3	22,590	46.4	18,117	46.2	4,473	0.78
NT-proBNP	74.0	36,092	73.9	28,863	74.7	7,229	0.11
BNP	83.1	40,489	82.7	32,322	84.4	8,167	<0.001

^Chi square test between development and validation sets.

*****n=48,745 total reference hospitalizations in the entire dataset.

**n=39,066 reference hospitalizations in the development set.

***n=9,679 reference hospitalizations in the validation set.

Table 4.8. Prediction metrics for the LaCE score.

Predictive Threshold*	TP	FN	FP	TN	PPV (post-test probability**)	LR+
9	2081	0	7782	0	0.21	1.00
10	2081	0	7782	0	0.21	1.00
11	2081	0	7781	1	0.21	1.00
13	2081	0	7780	2	0.21	1.00
14	2081	0	7779	3	0.21	1.00
15	2081	0	7778	4	0.21	1.00
16	2081	0	7775	7	0.21	1.00
17	2081	0	7773	9	0.21	1.00
18	2080	1	7773	9	0.21	1.00
19	2080	1	7771	11	0.21	1.00
20	2078	3	7770	12	0.21	1.00
21	2078	3	7767	15	0.21	1.00
22	2078	3	7764	18	0.21	1.00
23	2078	3	7762	20	0.21	1.00
24	2078	3	7760	22	0.21	1.00
25	2078	3	7754	28	0.21	1.00
26	2078	3	7747	35	0.21	1.00
27	2077	4	7742	40	0.21	1.00
28	2077	4	7741	41	0.21	1.00
29	2076	5	7738	44	0.21	1.00
30	2075	6	7731	51	0.21	1.00
31	2075	6	7730	52	0.21	1.00
32	2075	6	7727	55	0.21	1.00
33	2075	6	7727	55	0.21	1.00
34	2075	6	7720	62	0.21	1.01
35	2075	6	7717	65	0.21	1.01
36	2075	6	7711	71	0.21	1.01
37	2074	7	7704	78	0.21	1.01
38	2074	7	7696	86	0.21	1.01
39	2073	8	7685	97	0.21	1.01
40	2070	11	7661	121	0.21	1.01
41	2067	14	7636	146	0.21	1.01
42	2064	17	7609	173	0.21	1.01
43	2057	24	7567	215	0.21	1.02
44	2050	31	7504	278	0.21	1.02

45	2042	39	7444	338	0.22	1.03
46	2032	49	7384	398	0.22	1.03
47	2024	57	7275	507	0.22	1.04
48	2005	76	7196	586	0.22	1.04
49	1984	97	7084	698	0.22	1.05
50	1970	111	6984	798	0.22	1.05
51	1946	135	6803	979	0.22	1.07
52	1924	157	6636	1146	0.22	1.08
53	1902	179	6491	1291	0.23	1.10
54	1864	217	6286	1496	0.23	1.11
55	1834	247	6136	1646	0.23	1.12
56	1776	305	5844	1938	0.23	1.14
57	1735	346	5626	2156	0.24	1.15
58	1629	452	5182	2600	0.24	1.18
59	1529	552	4800	2982	0.24	1.19
60	1383	698	4332	3450	0.24	1.19
61	1214	867	3784	3998	0.24	1.20
62	1033	1048	3277	4505	0.24	1.18
63	821	1260	2548	5234	0.24	1.20
64	561	1520	1756	6026	0.24	1.19
65	398	1683	1285	6497	0.24	1.16
66	170	1911	560	7222	0.23	1.14

*LaCE score

**Compared to a pre-test probability of 20.9% using prevalence

TP: true positives; FN: false negatives; FP: false positives; TN: true negatives

PPV: positive predictive value; LR+: positive likelihood ratio

(--): no value

Table 4.9. AUROCs for various combinations of feature groupings for the XGBoost classifier.

Features	No. of predictors	AUC
ALL Features	171	0.651
All features without labs	156	0.645
Admission Characteristics + Healthcare Utilization + Comorbidity + Frailty + ACSC + Labs	107	0.649
Demography +Admission Characteristics +Healthcare Utilization + Comorbidity + Frailty + ACSC + Labs	115	0.648
Admission Characteristics +Healthcare Utilization + Comorbidity + Frailty + ACSC	75	0.633
Healthcare Utilization	8	0.618
Labs	32	0.595
Demography + Admission Characteristics + Drug Utilization	47	0.579
Comorbidity + Frailty	45	0.579
Demography + Admission Characteristics + Drug Utilization + Drug Adherence	56	0.578
Comorbidity	32	0.578
Admission Characteristics	15	0.574
Admission Characteristics + Cardiac Procedures	38	0.572
Demography + Admission Characteristics	23	0.566
Drug Utilization	24	0.558
Drug Utilization + Drug Adherence	33	0.550
Frailty	13	0.548
ACSC	7	0.545
Drug Adherence	9	0.521
Cardiac Procedures	23	0.515
Demography	8	0.507

Note: Frailty is a sub-group under comorbidities; ACSC: ambulatory care sensitive conditions

Table 4.10. Sub-group analysis stratified by type of heart failure.

Sub-group	Number of patients (%)	Number of hospitalizations (%)	c-statistic**
HFrEF	2,918 (29.64)	12,684 (26.02)	0.67
HFmEF	1,225 (12.44)	6,016 (12.34)	0.61
HFpEF	5,702 (57.92)	30,045 (61.64)	0.63
HF specific reference hospitalizations*	3,977 (n/a)	7,574 (n/a)	0.60

Note: heart failure with reduced ejection fraction (HFrEF); heart failure with mid-range ejection fraction (HFmEF); heart failure with preserved ejection fraction (HFpEF); n/a: not applicable

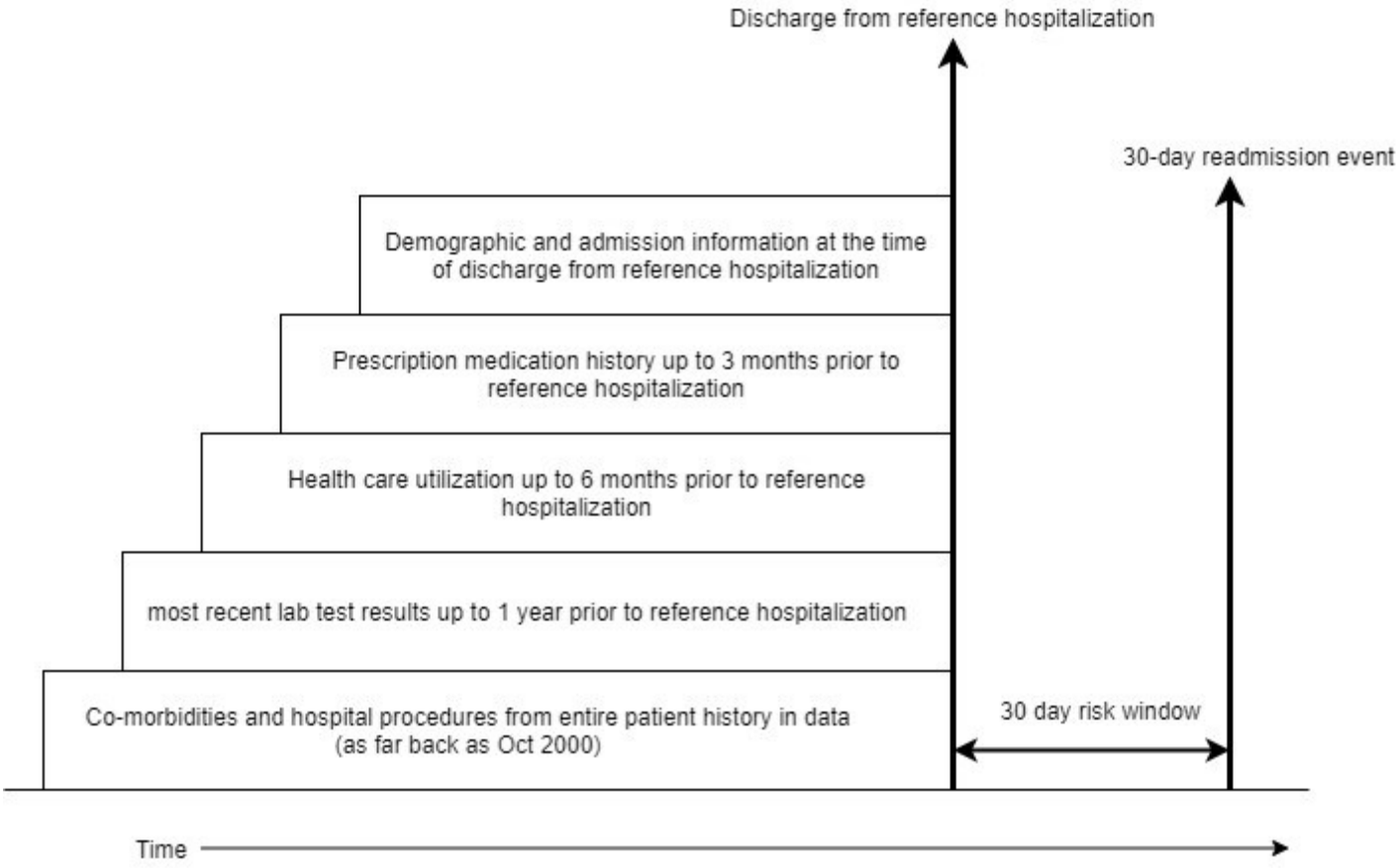
*Determined using main diagnosis field and ICD-10 code I50

**Discrimination performance done on validation set

Table 4.11. Distribution of missing laboratory data in the HF specific reference hospitalization subset.

	% Missing	Number Missing (total n=7,574)
Blood Urea Nitrogen	20.6	1,559
MCV	18.5	1,400
Hemoglobin	18.5	1,400
Total Cholesterol	48.2	3,653
Potassium	18.4	1,398
Creatinine	18.4	1,398
eGFR	18.6	1,413
Cholesterol Ratio	55.7	4,221
Sodium	18.4	1,398
Hematocrit	18.5	1,400
ALT	24.0	1,814
Ferritin	46.1	3,493
Albumin	31.1	2,354
A1C	40.9	3,100
NT-proBNP	52.1	3,945
BNP	75.3	5,703

Figure 4.5. Timelines of data capture for candidate predictor categories.



Note: Reference hospitalization from April 2012-March 2019 were used to train Machine Learning models.

Figure 4.6. Frequency of patients with multiple admissions.

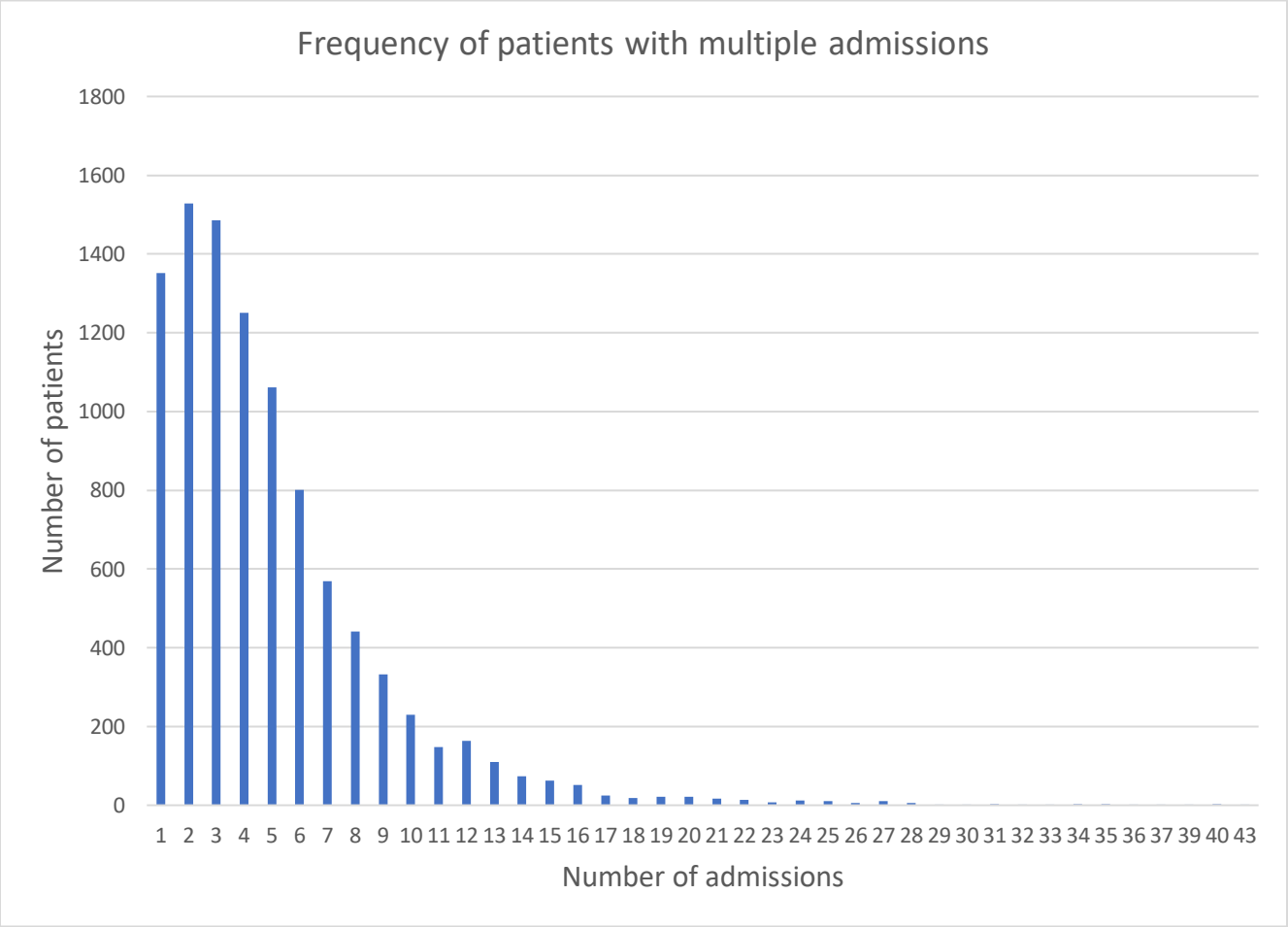
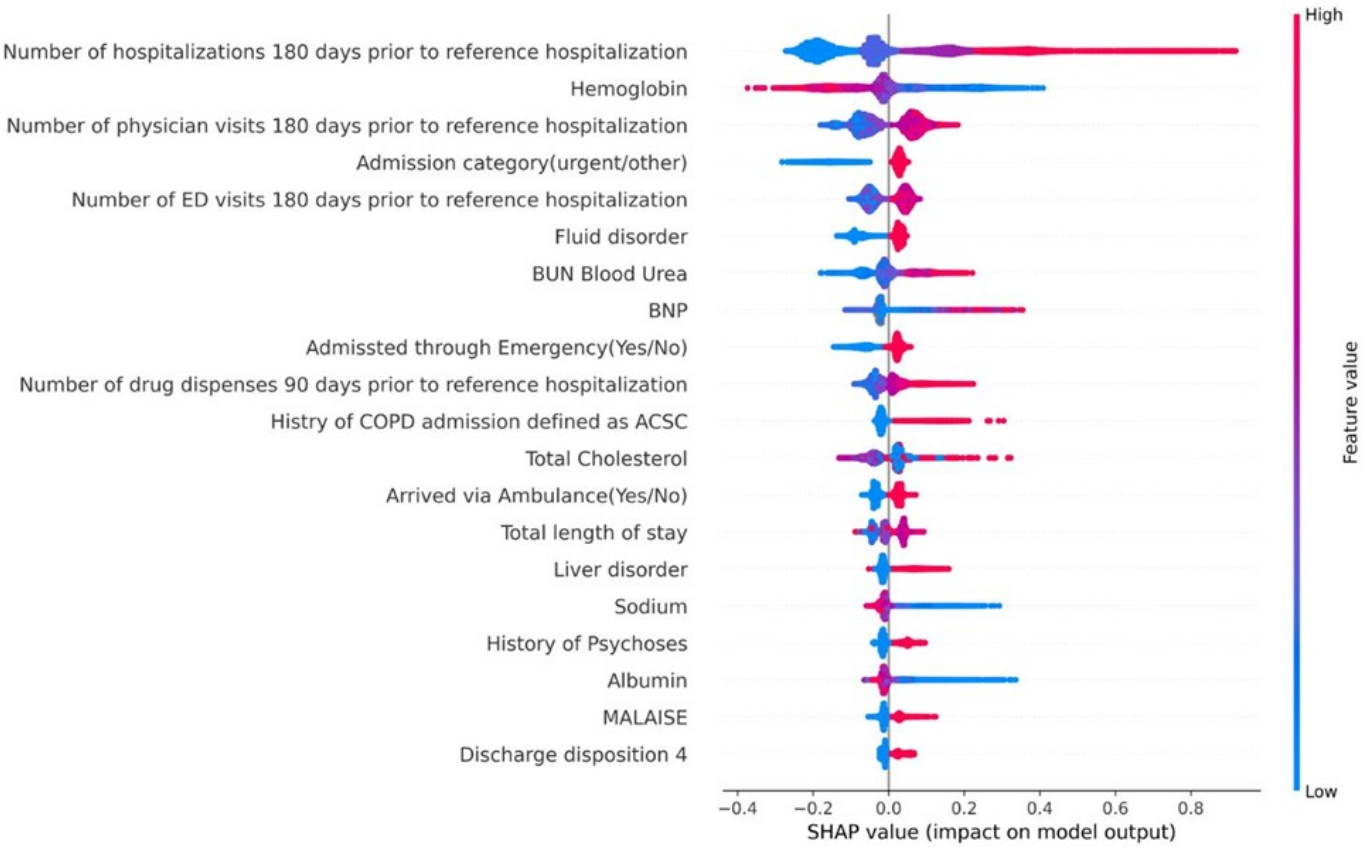


Figure 4.7. SHAP plot indicating influence of various predictors on risk of readmission*



*These predictors are rank-ordered according to feature importance with importance decreasing from top to bottom (See Figure 4.4 in main text); variable impact does not have a causal nor statistical meaning but is simply a measure of the influence a predictor has on the ML model output. Red indicates higher values of the predictors while blue indicates lower values and lines to the right of the 0.0 on the x-axis are associated with readmissions while those to the left are less associated with readmissions. For example, lower levels of hemoglobin (blue) are predictive of readmissions (blue is to the right of the 0.0 on x-axis) where in some instances, lower hemoglobin (blue) has a higher influence on readmissions (further to the right of the 0.0 on the x-axis). Similarly, for binary variables, a history of liver disease or peptic ulcer disease (red colour and to the right of the 0.0 on the x-axis) is predictive of readmissions to varying extents.

References for Chapter 4

4. Jencks SF, Williams MV, Coleman EA. Rehospitalizations among Patients in the Medicare Fee-for-Service Program. *New England Journal of Medicine.* 2009;360(14):1418-1428.
5. Alberta Health Services. *AHS Report on Performance FY 2017-18. Unplanned Medical Readmissions.* 2018.
9. Canadian Institute for Health Information. All-Cause Readmission to Acute Care and Return to the Emergency Department. 2012.
11. 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. Alberta Machine Intelligence Institute. Machine Learning Process Lifecycle. In:2019.
15. Shah NH, Milstein A, Bagley P, Steven C. Making Machine Learning Models Clinically Useful. *JAMA.* 2019;322(14):1351-1352.
16. Morgenstern JD, Buajitti E, O'Neill M, et al. Predicting population health with machine learning: a scoping review. *BMJ Open.* 2020;10(10):e037860.
17. 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.
18. Luo W, Phung D, Tran T, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res.* 2016;18(12):e323.
19. 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.
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.
22. 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. Frankl SE, Breeling JL, Goldman L. Preventability of emergent hospital readmission. *The American journal of medicine.* 1991;90(6):667-674.
32. Cognilytica. Cognitive Project Management for Artifical Intelligence Methodology. In:2020.

34. Mooney SJ, Pejaver V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annual Review of Public Health.* 2018;39(1):95-112.

37. 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.

39. Rose S. Machine Learning for Prediction in Electronic Health Data. *JAMA Network Open.* 2018;1(4):e181404-e181404.

42. Frizzell JD, Liang L, Schulte PJ, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA cardiology.* 2017;2(2):204-209.

45. Donders ART, Van Der Heijden GJ, Stijnen T, Moons KG. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology.* 2006;59(10):1087-1091.

46. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. *Journal of Clinical Epidemiology.*

50. 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.

51. Hu Z, Melton GB, Arsoniadis EG, Wang Y, Kwaan MR, Simon GJ. Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *Journal of Biomedical Informatics.* 2017;68:112-120.

52. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological methods.* 2002;7(2):147.

53. Sharafoddini A, Dubin JA, Maslove DM, Lee J. A new insight into missing data in intensive care unit patient profiles: Observational study. *JMIR medical informatics.* 2019;7(1):e11605.

56. Talhouk Aline. AI Predictive Analytics: pathways from research to the clinic. Paper presented at: AI Enabled Care: Building Collaboration for Deeper Learning and Better Care; April 2021, 2021; Michener Institute.

59. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA.* 2018;320(21):2199-2200.

66. 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.

68. 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.

77. World health Organization. Classification of Diseases (ICD). 2019; https://www.who.int/classifications/icd/icdonlineversions/en/. Accessed Jun 2020.

87. 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.

91. Molnar C. *Interpretable machine learning. A Guide for Making Black Box Models Explainable.* 2019.

92. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Paper presented at: Advances in neural information processing systems2017.

94. 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.

95. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Paper presented at: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining2016.

96. The pandas development team. pandas-dev/pandas: Pandas. 2020; https://doi.org/10.5281/zenodo.3509134, Jan 2021.

97. Ezekowitz JA, O'Meara E, McDonald MA, et al. 2017 Comprehensive Update of the Canadian Cardiovascular Society Guidelines for the Management of Heart Failure. *Canadian Journal of Cardiology.* 2017;33(11):1342-1433.

98. Yam CH, Wong EL, Chan FW, et al. Avoidable readmission in Hong Kong-system, clinician, patient or social factor? *BMC health services research.* 2010;10(1):311.

99. Alberta Health. Performance Measure Definition-30 day Overall Readmission Rate. 2014; https://open.alberta.ca/dataset/c7e3fc16-7aea-455c-96a1-20811a640b1a/resource/63ee45db-a066-4298-b63d-ba254eee5dc5/download/PMD-30-Day-Readmission-Rate.pdf.

100. Feltner C, Jones CD, Cené CW, et al. Transitional Care Interventions to Prevent Readmissions for Persons With Heart Failure. *Annals of Internal Medicine.* 2014;160(11):774-784.

101. van Walraven C, Dhalla IA, Bell C, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Canadian Medical Association Journal.* 2010;182(6):551-557.

102. Au AG, McAlister FA, Bakal JA, Ezekowitz J, Kaul P, van Walraven C. Predicting the risk of unplanned readmission or death within 30 days of discharge after a heart failure hospitalization. *American Heart Journal.* 2012;164(3):365-372.
103. Wang H, Robinson RD, Johnson C, et al. Using the LACE index to predict hospital readmissions in congestive heart failure patients. *BMC Cardiovascular Disorders.* 2014;14(1):97.
104. Yazdan-Ashoori P, Lee SF, Ibrahim Q, Van Spall HG. Utility of the LACE index at the bedside in predicting 30-day readmission or death in patients hospitalized with heart failure. *American heart journal.* 2016;179:51-58.
105. Mortazavi BJ, Downing NS, Bucholz EM, et al. Analysis of machine learning techniques for heart failure readmissions. *Circulation: Cardiovascular Quality and Outcomes.* 2016;9(6):629-640.
106. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology.* 2019;110:12-22.
107. Shin S, Austin PC, Ross HJ, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC Heart Failure.* 2021;8(1):106-115.
108. Ghimire A, Fine N, Ezekowitz JA, Howlett J, Youngson E, McAlister FA. Frequency, predictors, and prognosis of ejection fraction improvement in heart failure: an echocardiogram-based registry study. *European heart journal.* 2019;40(26):2110-2117.
109. Alberta College of Pharmacy. 2019; https://abpharmacy.ca/. Accessed Sept 2019.
110. Canadian Institute for Health Information. 2019; https://www.cihi.ca/en.
111. van Walraven C, Wong J, Forster AJ. LACE+ index: extension of a validated index to predict early death or urgent readmission after hospital discharge using administrative data. *Open Medicine.* 2012;6(3):e80.
112. Keyko K JL. *Pathway Pearls LACE Index Scoring.* Alberta Health Services; June 2018 2018.
113. Jiang W, Siddiqui S, Barnes S, et al. Readmission Risk Trajectories for Patients With Heart Failure Using a Dynamic Prediction Approach: Retrospective Study. *JMIR medical informatics.* 2019;7(4):e14756-e14756.
114. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Chow BJ, Dwivedi G. Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death. *PLOS ONE.* 2019;14(6):e0218760.

115. Leppin AL, Gionfriddo MR, Kessler M, et al. Preventing 30-day hospital readmissions: a systematic review and meta-analysis of randomized trials. *JAMA internal medicine.* 2014;174(7):1095-1107.
116. Vader JM, LaRue SJ, Stevens SR, et al. Timing and Causes of Readmission After Acute Heart Failure Hospitalization-Insights From the Heart Failure Network Trials. *J Card Fail.* 2016;22(11):875-883.
117. Reddy YN, Borlaug BA. Readmissions in Heart failure: It's more than just the medicine. Paper presented at: Mayo Clinic Proceedings2019.
118. Retrum JH, Boggs J, Hersh A, et al. Patient-identified factors related to heart failure readmissions. *Circ Cardiovasc Qual Outcomes.* 2013;6(2):171-177.
119. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research.* 2011;12:2825-2830.
120. XGBoost. Python API Reference. https://xgboost.readthedocs.io/en/latest/python/python_api.html#module-xgboost.sklearn. Accessed August 2021.
121. Ali M. PyCaret: An open source, low-code machine learning library in Python version 2.3. April 2020; https://pycaret.org/about.
122. Jin H, Song Q, Hu X. Auto-keras: An efficient neural architecture search system. Paper presented at: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining2019.
123. H20.ai. 2021; https://www.h2o.ai/company/.
124. Schmidhuber J, Hochreiter S. Long short-term memory. *Neural Comput.* 1997;9(8):1735-1780.
125. Ashfaq A, Sant'Anna A, Lingman M, Nowaczyk S. Readmission prediction using deep learning on electronic health records. *Journal of Biomedical Informatics.* 2019;97:103256.
126. Dharmarajan K, Hsieh AF, Lin Z, et al. Diagnoses and Timing of 30-Day Readmissions After Hospitalization for Heart Failure, Acute Myocardial Infarction, or Pneumonia. *JAMA.* 2013;309(4):355-363.
127. Alberta Health Services and Government of Alberta. Admissions for Ambulatory Care Sensitive Conditions Indicator Definition. 2011.
128. van Smeden M, Groenwold RHH, Moons KGM. A cautionary note on the use of the missing indicator method for handling missing data in prediction research. *Journal of Clinical Epidemiology.*

129. Pencina MJ, D'Agostino RB, Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in medicine.* 2011;30(1):11-21.
130. Mitnitski A, Howlett SE, Rockwood K. Heterogeneity of Human Aging and Its Assessment. *The Journals of Gerontology: Series A.* 2016;72(7):877-884.
131. Liu L-F. The Health Heterogeneity of and Health Care Utilization by the Elderly in Taiwan. *International Journal of Environmental Research and Public Health.* 2014;11(2):1384-1397.
132. Bayati M, Braverman M, Gillam M, et al. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PloS one.* 2014;9(10):e109264.
133. Komorowski M. Artificial intelligence in intensive care: are we there yet? *Intensive care medicine.* 2019;45(9):1298-1300.

Chapter 5: Predicting 30-day risk from benzodiazepine/Z-drug dispensations in older adults using administrative data: a prognostic machine learning approach

Objective: To develop a machine-learning (ML) model using administrative data to estimate risk of adverse outcomes within 30-days of a benzodiazepine (BZRA) dispensation in older adults for use by health departments/regulators.

Design, Setting and Participants: This study was conducted in Alberta, Canada during 2018-2019 in Albertans 65 years of age and older. Those with any history of malignancy or palliative care were excluded.

Exposure: Each BZRA dispensation served as the unit of analysis.

Main Outcomes and Measures: ML algorithms were developed on 2018 administrative data to predict risk of unplanned emergency department visit, hospitalization or death within 30-days of a BZRA dispense. Validation on 2019 administrative data was done using XGBoost to evaluate discrimination, calibration and other relevant metrics on ranked predictions. Daily and quarterly predictions were simulated on 2019 data.

Results: 65,063 study participants were included which represented 633,333 BZRA dispenses during 2018-2019. The validations set had 314,615 dispenses linked to 55,928 all-cause outcomes representing a pre-test probability of 17.8%. C-statistic for the XGBoost model was 0.75. Measuring all-cause risk at the end of 2019, the top 0.1 percentile of predicted risk had a LR+ of 40.31 translating to a post-test probability of 0.90. Daily and quarterly classification simulations resulted in uninformative predictions with LR+'s less than 10 in all risk prediction categories. Previous history of admissions was ranked highest in variable importance.

Conclusion: Developing ML models using only administrative health data may not provide health regulators with sufficient informative predictions to use as decision aids for potential interventions, especially if considering daily or quarterly classifications of BZRA risks in older adults. ML models may be informative for this context if yearly classifications are preferred. Health regulators should have access to other types of data to improve ML prediction.

Introduction

Health jurisdictions like Alberta Health have prioritized reducing avoidable inpatient admissions and associated costs; Alberta, Canada has higher rates of hospitalizations compared to other provinces[134]. A specific area of concern is the wide use of benzodiazepine receptor modulators (BZRAs) in older adults. Canadian clinical practice guidelines suggest that BZRA treatment is appropriate for short-term use in adults aged 20 to 64 and in some cases as second-line treatment [135-137]. Use of BZRAs outside of these recommendations is considered "potentially inappropriate" given the potential for adverse effects, especially in older adults[135,138,139]. In fact, BZRAs are not recommended at all (regardless of duration) in older persons as first line therapy for insomnia[136]. A 2006 study in British Columbia, Canada found that 3.5% of the population were considered long-term users of BZRAs and 47% were over the age of 65[26]. Furthermore, a recent study reported that 10% of Albertans in 2015 received a BZRA with the prevalence of use increasing with age[140].

Despite warnings from experts and regulatory bodies, BZRAs continue to be prescribed to older adults at alarmingly high rates leading to adverse outcomes such as hospitalizations. Indeed, the risk of hospitalizations or death can more than double in older adults taking BZRAs[28]. The main strategies to address BZRA use involve safe drug-use guidelines and stewardship efforts from health regulators such as the Tracked Prescription Program (TPP) at the College of Physicians and Surgeons of Alberta (CPSA)[141]. The American Geriatrics Society also maintains its Beers Criteria ® for potentially inappropriate medication use in older adults in which BZRAs are not recommended due to the risks of cognitive impairment, falls, fractures and accidents[142]. Furthermore, other groups have targeted BZRAs in deprescribing initiatives, especially in older adults[136,143]. However, none of these measures involve risk prediction at the individual patient level using absolute probabilities.

Given these potentially preventable outcomes, there is substantial interest from health regulators and systems[134] to predict inpatient admissions and newer prediction tools may address this interest. Individual risk from BZRA dispensations is an important clinical outcome in which machine learning (ML) prediction can play a role. ML prediction offers a way to quantify individual risk within a well-defined project framework emphasizing deployment into

real world settings. Furthermore, ML approaches are gaining traction in the clinical prediction space[16,17].

Supervised machine learning (ML)[11,15] is an approach that uses computer algorithms to build predictive models in the clinical setting that can make use of large amounts of available administrative data[12,13], all within a well-defined process[14]. Supervised ML trains on labelled data to develop prediction models that are specific to different populations and, in many cases, can provide better predictive performance than traditional statistical models[13,37,79] based on sampling populations[29].

To our knowledge, there are no risk prediction tools in use which can quantify the risk of BZRA dispensations at the individual level, something that health regulators like the CPSA are interested in. The objective of our study is to predict the 30-day risk of an adverse event pursuant to a BZRA dispensation in adults over 65 years of age in Alberta using supervised ML methods trained on readily available administrative data. We will develop and validate commonly used ML algorithms and evaluate prediction performance and utility using routine clinical prediction metrics[18-20,38]. Our analysis will also include ML prediction simulations. This study will provide Alberta Health, and health jurisdictions in general, with analytic options to assess the value and implementation of ML classifiers.

Methods

Study Design, Setting, and Participants

This prognostic study used a supervised ML approach which trained ML models on administrative health data in Alberta, Canada during 2018-2019. Albertans 65 years of age and older who were dispensed a BZRA during the study period were included in this study. Those with any history of malignancy or palliative care were excluded.

Data Sources

To develop our ML model, we linked various administrative health data sets available in Alberta, Canada using unique anonymized patient identifiers to establish a thorough description of demographics, medical history, drug exposures and outcomes. The following databases were linked: *1) Pharmaceutical Information Network (PIN):* data on all dispensing

records from community pharmacies irrespective of coverage status and according to the guidelines from the Alberta College Pharmacy[109]; Anatomical Therapeutic Chemical classification (ATC) codes[80] were used to identify BZRA dispensations and their respective molecules (Table 5.4), *2) Population and Vital Statistics Data (VS, Alberta Services)*: sex, age, date of birth, death date, immigration and emigration data within the province, and underlying cause of death according to the World Health Organization algorithm using ICD codes (International Statistical Classification of Diseases and Related Health Problems[77]), *3) Hospitalizations and Emergency Department Visits (NACRS [National Ambulatory Care Reporting System], DAD [Discharge Abstract Database]):* all services, length of stay, diagnosis (up to 25 ICD-10 based diagnoses). Data and coding accuracy are routinely validated both provincially and centrally via the Canadian Institute for Health Information (CIHI)[110], 4*) Physician Visits/Claims (Alberta Health):* date of service, up to three ICD codes associated with the claim, procedure and billing information, and *5) Provincial Laboratory, Alberta Health Services:* all laboratory services conducted within the hospital or community.

These linked databases represent a labelled dataset used to develop and validate ML models as each instance of a BZRA dispensation was labelled with an outcome. Our data spanned the period 2013-2019.

Measures and Outcomes

For this study, the unit of analysis was at the BZRA dispensation level such that each dispensation served as a potential instance to predict our outcome. We chose this level of analysis to be consistent with others[37], to have more data to train the ML model, and to accurately represent use in the real world in which health regulators may want to assess the risk for each instance rather than a single or random dispensation.

Our outcome was an all-cause unplanned hospitalization, emergency department (ED) visit or death within 30-days of a BZRA dispensation. For comparison, we also defined a cause specific composite outcome of accidents, poisonings, falls and injuries, all of which are recognized as adverse outcomes from BZRAs in older adults (Table 5.5). 30-day risk windows

are commonly used by health systems for risk assessments[82]. Follow-up and predictions started after each BZRA dispensation.

Predictors and ML Methods

All candidate predictors, both derived and directly pulled from the data, were obtained from the linked datasets (Table 5.4). The feature categories included demographics (age, sex), drug utilization (ATC codes, oral morphine equivalents[88], concurrent use of BZRAs and opioids, number of dispensations, number of unique molecules) and health care utilization (prior admissions, number of unique prescribers and pharmacies). Co-morbidity history consisted of Elixhauser score categories[87] (commonly used for risk adjusting) and other conditions[144]. Routine lab measurements were also incorporated and were kept as continuous variables thus, were not categorized to minimize loss of variation. Time sensitive variables were derived and represented in the drug and health care utilization categories. Depending on the predictor, we used data from 30 days to 5 years prior to the BZRA dispensation to generate model features (Figure 5.5).

We used commonly available ML algorithms and approaches to train our ML models[11,14,17,50]. These included logistic regression, Gradient Boosting, Linear Support Vector, Multi-layer Perceptron, XGBoost and AdaBoost. Our outcome labelled dataset was split such that BZRA dispenses in 2018 comprised the development set for ML model training and hyperparameter tuning while those in 2019 comprised the validation set in which ML model performance was assessed. We created two validation sets, one which was not a true external independent set such that participants in the development set could also be represented in the validation set and another which only included new BZRA patients in 2019 thus creating an independent validation set (Figure 5.1). Both validation sets were created to represent real world use scenarios in which ML models are trained on the entire population they will be deployed in (patients could be repeatedly encountered) and to evaluate ML prediction performance in new, out of sample patients.

Based on results from our preliminary work, we performed ML model evaluation using the XGBoost classifier because it performed better than others and is considered "explainable" ML. We did however report discrimination performance of the other ML models.

We performed k=5 cross fold validation in the development set to tune hyperparameters. With XGBoost, we tuned for tree height, and number of trees.[95,120]. During hyperparameter tuning of XGBoost, we explored the use of scale_pos_weight[119] to address class imbalance, however, we did not include it in the XGBoost model development to improve calibration.

Missing Data

We anticipated missing data when handling laboratory results. It should be noted that the Provincial Laboratory fully captures all performed laboratory tests and that any "missing" data are laboratory tests that were not ordered as opposed to missing. Missing laboratory data is classified as either missing at random or missing not at random[45,50] and imputation methods are not favourable with ML prediction tools intended for deployment[52]. Instead, we included missing indicator variables as others have done[46,51,53] although we understand this is controversial[128]. However, if the models are to be deployed in real-world settings, the missing indicator approach is most practical. Tree-based ML algorithms (e.g. XGBoost) are able to handle missing data better than regression based algorithms[50]. We made use of the *Sparsity-aware split finding* feature[95] when training the XGBoost algorithm to address missing data without the need for missing indicators. All other ML algorithms required the use of missing indicators. Other ML studies excluded missing medical data or used imputation to handle missing data, both recognized as leading to high bias[42,107,113].

Analysis and Prediction Evaluation

We first described co-morbidity characteristics in the development vs validation groups (entire 2019 validation set) and between those who did and did not experience the outcome using chi-square tests and t-tests. This descriptive analysis was done at the patient level in which each participant could be represented by multiple instances of BZRA dispensations. Outcome event rates were also reported. The final piece of our descriptive analysis was

describing the distribution of missing lab data. We did this for 2018, 2019 individually and for the combination of 2018-2019 data.

The validation sets were used to evaluate ML model prediction performance using metrics commonly applied to clinical prediction models[18-20,38]. Mentioned previously, our validation assessments were done using XGBoost. As is done in many ML prediction studies[17], we assessed our ML model discrimination performance by estimating the area under the receiver operating characteristics curve (AUROC). For binary classification studies like ours, AUROC curves correspond to c-statistics which are a measure of model discrimination performance, the extent to which a model predicts a higher probability of an outcome among instances that actually had the outcome compared to those that did not[19]. We reported AUROCs for the all-cause outcome using all of the ML algorithms we trained. We also included AUROCs of the XGBoost model using fewer feature categories for all-cause outcomes, specifically the one that excluded lab test results because of the anticipated data missingness in this category. For comparison, we reported AUROCs on the cause specific composite outcome and for individual outcomes which comprised the composite outcome. Precision-recall curves (PRC) were also included[89] using the all-cause outcome.

We provided a calibration plot[19] for the XGBoost model using both validation sets; calibration is considered an important property of any prediction model and reflects the extent to which predicted values align with observed values and is most often illustrated by a plot of observed vs predicted[18,19]. Calibration was done on both validation sets and within the subgroup of sex. As well, we added a negative predictive value (NPV) vs. predicted risk plot to highlight the relationship between low predicted risk and true negatives (those who did not experience the outcome).

We then reported two methods for assessing the clinical utility of the ML model. The first involved ranking our predicted risks, as others have done[66], by categorizing them into percentiles (e.g., deciles) or keeping them in absolute numbers (e.g., top 10 highest risk dispenses). At each of these category cut-off points, we reported prediction performance metrics. These included positive likelihood ratios (LR+)[20,145], true/false positives, true/false

negatives, and positive predictive values (PPV, equivalent to post-test probability). These metrics were also reported on the actual thresholds of predicted risk outputted by the ML classifier. We carried out this analysis on both validation sets and measured these metrics at the end of 2019.

In the second method, we performed a decision curve analysis[21] in which the net benefit of our ML model is compared against two alternatives, namely, intervening on all BZRA dispensations or on none, using the entire range of probability threshold cut-off points. This comparison is done by using predicted probabilities from our ML model and comparing them against a probability threshold to aid a decision. Again, we did this on both validation sets using all-cause outcomes. Thus, if a health regulator is interested in intervening, for example, on the top 1 percentile of predicted risk or top 10 highest risk predicted BZRA dispenses, then method 1 could be considered. Alternatively, if BZRA dispensations above a certain predicted risk threshold are of interest, then method 2 could be informative. Either way, the amount of workload created by identifying high risk dispenses is an important factor for ML deployment.

We simulated all-cause outcome predictions to view the capabilities of the XGBoost model if deployed into a workflow. These included predictions measured daily and quarterly which progressively excluded participants once they were already flagged as high risk. Filtering out previously flagged patients represents a more realistic scenario for health regulators as it is not practical to repeatedly identify the same high-risk patients. For this simulation, previously flagged participants were excluded for the entire year keeping in mind that a health regulator could exclude on a monthly or quarterly basis. Specifically, we simulated on 2019 data by categorizing the highest predicted BZRA dispenses (e.g., top 10 highest risk BZRA dispenses) and then measuring the same metrics described above for each category. We also simulated the number of 30-day events per 100 daily dispenses stratified by percentiles of risk.

Because ML models do not estimate an interpretable quantity relating predictors to outcomes, it is not appropriate to summarize that relationship with a single parameter. Instead, the impact of individual predictors can be summarized using “variable importance”, which is a rank-ordering of variables that are most important for the ML model’s prediction

performance[50]; variable importance does not have a causal or statistical meaning. To address interpretability[59] of our ML model, we reported feature importance[50] and feature impact using SHAP (Shapley Additive Explanations) value plots[91,92], which would give health regulators some insight on how the ML model was influenced in its predictions.

This study followed the TRIPOD[68] and other guidelines[11,17] specific to ML projects. All analyses were done using Python (version 3.6.8, Python Software Foundation), SciKit Learn[94] (version 0.23.2), SHAP[92] (version 0.35), XGBoost[95] (version 0.90), Pandas[96] (version 1.0.5) and STATA/MP V.15.1 (StataCorp). This study received ethics approval from the University of Alberta ethics board (Pro00083807_AME6).

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

A total of 65,063 participants were included in this study representing 633,333 BZRA dispenses during 2018-2019. During this time, there were 114,299 (18.0%) all cause outcomes that were linked to BZRA dispenses within 30 days. In comparison, the cause specific composite outcome occurred at a much lower rate at 1.8% (n=5,998 for 2018 and n=5,712 for 2019). Dispenses in 2018 and 2019 comprised the development and validation sets, respectively (Figure 5.1). The full validation set had 314,615 BZRA dispenses while the independent (out of sample) validation set had 37,070 BZRA dispenses. We measured 55,928 all cause outcomes in the full validation set representing a pre-test probability of 17.8% while the corresponding numbers in the independent validation set were and 7,324 and 19.7%, respectively. Characteristics were different between those who did and did not experience the all-cause outcome (Table 5.6), as expected, and were more similar between those in the development and validation sets (Table 5.7).

The number of instances without a laboratory test result ranged from 15-64% of BZRA dispenses (Table 5.8).

C-statistics ranged from 0.67 for logistic regression to 0.75 for XGBoost (Table 5.9). Further assessing the XGBoost model, the c-statistic when using all of the datasets was 0.75 compared to 0.66 when using only demographic and drug history features (Table 5.10). Of note, the c-statistic for the XGBoost model developed without the lab test features was 0.75 indicating that excluding lab features did not influence discrimination performance. The c-statistic for the out of sample validation set was 0.74, similar to the entire validation set (Figure 5.6). For comparison, the c-statistic for the cause specific composite outcome was 0.69 (Table 5.10). Both validation sets had similar PRCs (Figure 5.6).

For both validation sets, the calibration plots indicate that our predictions were marginally aligned with the observed event rates with slight overestimation of risk and that most BZRA dispenses were classified as lower risk (Figure 5.7). Calibration was similar across sex in both validation sets (Figure 5.8). Lower predicted risks were accompanied by fewer actual outcomes signaling higher NPVs at lower predicted risks (Figure 5.9).

After we ranked and grouped our predicted risks at the end of 2019, the categories with the highest risk predictions expectedly had the highest PPVs (post-test probabilities) and LR+'s. The top 0.1 percentile of predicted risk had LR+ of 40.31 which translated to a PPV of 0.90 using the entire validation set. The corresponding numbers for the out of sample validation set were 134.03 and 0.97, respectively (Table 5.1). Similar results were observed in the top 20 highest risk dispenses with a LR+ and PPV of 48.57 and 0.91, respectively (Table 5.11). There was also an increase in LR+ and PPV as the threshold of predicted risk increased (Table 5.12).

When we performed the decision curve analysis across the entire range of predicted probability thresholds, the XGBoost classifier provided some additional value over treating all or none with a potential intervention within the range of 0.1-0.5 of predicted probability (Figure 5.2).

In our simulations, predictions classified quarterly were more informative than those done daily. The top 10 daily highest risk BZRA dispenses had a LR+ of 5.60 (Table 5.2) while the

corresponding LR+ for the quarterly top 500 ranged from 5.71-9.83 (Table 5.3). Classifying the top 100 daily highest risk dispenses or top 10,000 done quarterly produced LR+'s of 2.39 and 1.29-2.64, respectively. When we simulated based on top percentiles of predicted risks and events per 100 BZRA dispenses, the top 1, 5 and 10 percentiles of predicted risk had higher event rates than the baseline risk, although there was noticeable overlap between the top percentiles (Figure 5.3).

With respect to ML interpretability, previous history of admissions was ranked highest in variable importance (Figure 5.4) with a positive history being suggestive of a higher risk (Figure 5.10), which is expected.

Discussion

In this study, we developed ML models to predict adverse outcomes pursuant to a BZRA dispensation and further explored the XGBoost classifier for validation. We presented two analytic options for health regulators to consider for implementing ML decision support, namely acting on the highest ranked predictions or on probability threshold cut-off points. Discrimination, calibration and net benefit analysis are important aspects in determining the clinical utility of a prediction model[21]. When predictions were classified at the end of 2019, our XGBoost model displayed strong discrimination and calibration performance in both validation sets signalling similar prediction performance in both previous and new BZRA patients. The net benefit analysis showed that the ML model could provide some clinical utility over a small range of predicted probability thresholds; however, it may not be sufficient to deploy within real-world health systems.

Despite the strong prediction performance of our ML model when risk was measured at the end of 2019, the daily and quarterly simulations suggest otherwise; the frequency of prediction classification influences prediction performance metrics. Our predictions classified at yearend reported very informative LR+s (up to 134) while those in our daily-quarterly simulations did not surpass 10. LR+'s greater than 10 are considered strongly informative with conclusive changes from pre-test to post-test probabilities[20]. Health regulators should strongly consider this finding when deciding if daily, quarterly, or yearly classifications of risk are to be

implemented; yearly classification of risk may be of limited value. This finding could be partly explained by the limited capabilities of training ML models only on administrative data. It is well known that having additional types of data could better leverage ML prediction capabilities[16,34] thus making the case to increase data access and permissions for health regulators.

Our simulation results, which represent a more realistic use case scenario, indicate that predicting adverse events after a BZRA dispensation is a difficult undertaking using ML methods. The uninformative prediction performance reflects this difficulty which could be explained by data quality. It is estimated that more than 80% of the work done in ML projects is composed of data preparation[32]; there was not enough variation in our administrative datasets for informative ML prediction done daily or quarterly. To augment ML performance, our data could be linked to social factors data, which are known determinants of health however, this analysis cannot be done in most jurisdictions, including Alberta, Canada. The elderly are a very heterogeneous segment of society with regard to health status[130,131]. Indeed, incorporating the entire taxonomy of 'big' health data would likely improve ML prediction by linking biological, geospatial, electronic health records, personal monitoring and effluent sources of data all of which contribute to patient heterogeneity[34].

To date, we are unaware of any studies that assessed BZRA risk using ML methods like we have in Alberta, Canada or elsewhere. Our study benefited from complete records of hospitalizations, ED visits, physician claims and medication histories from anywhere in the province of Alberta. Furthermore, our analysis included informative metrics not commonly reported in clinical ML prediction studies. We also reported SHAP values to provide a measure of interpretability of our ML model. Another benefit is that by variably ranking the highest predicted BZRA dispenses, health regulators can control or adjust their workload to align with their capacity.

Limitations are mainly due to data issues. We did have a substantial number of dispenses without lab test data, however, excluding lab tests from ML training did not reduce the discrimination performance, a finding reported by others[51]. Although we incorporated

many different predictors derived from administrative datasets, we were not able to measure many other predictors that would substantially contribute to adverse BZRA outcomes, namely social factors. Others have pointed this limitation in their work as well[16,42] and that even with social factors data, ML prediction performance is not guaranteed but some improvement could be expected[118]. Our ML model was developed using data from Alberta, Canada and may not be generalizable to other jurisdictions. Nevertheless, the ML process makes it simple to develop and validate new models using local population specific data.

This study considered the perspective of health regulators and their mandate to monitor adverse outcomes from BZRA dispensations in older adults. Using ML developed on administrative health data alone may not provide informative predictions, especially if daily or quarterly reporting is desired. As reported in our analysis, there may be some benefit if yearly classifications are used. Although ML prediction may be promising, health regulators may require additional sources of data before implementing ML prediction as a decision aid leading to interventions; many jurisdictions do not permit the widespread use and sharing of administrative data. Having access to other types of data may improve ML performance and usefulness and should be explored. Furthermore, whether or not ML prediction and subsequent interventions can reduce adverse outcomes related to BZRAs is an area for future research.

Table 5.1. All-cause outcome prediction metrics stratified by top percentile of risk for the XGBoost classifier measured at the end of 2019 on both validation sets.

	Top percentile of risk	Threshold	TP	FN	FP	TN	PPV*	NPV	Se	Sp	LR+
Entire validation set	**0.01**	0.978	28	55,900	2	258,685	0.933	0.822	0.000501	0.999992	64.76
	0.1	0.947	244	55,684	28	258,659	0.897	0.823	0.004363	0.999892	40.31
	1	0.694	2,140	53,788	975	257,712	0.687	0.827	0.038263	0.996231	10.15
	5	0.498	7,988	47,940	6,396	252,291	0.555	0.840	0.142826	0.975275	5.78
	10	0.420	13,738	42,190	14,682	244,005	0.483	0.853	0.245637	0.943244	4.33
	25	0.306	27,326	28,602	43,999	214,688	0.383	0.882	0.488592	0.829914	2.87
	50	0.182	43,086	12,842	102,530	156,157	0.296	0.924	0.770383	0.603652	1.94
	75	0.084	51,077	4,851	170,417	88,270	0.231	0.948	0.913263	0.341223	1.39
	90	0.058	54,066	1,862	215,770	42,917	0.200	0.958	0.966707	0.165903	1.16
Independent validation set	**0.01**	0.973	4	7,320	0	29,746	1.000	0.803	0.000546	1.000000	(--)
	0.1	0.951	33	7,291	1	29,745	0.971	0.803	0.004506	0.999966	134.03
	1	0.716	242	7,082	123	29,623	0.663	0.807	0.033042	0.995865	7.99
	5	0.534	1,002	6,322	757	28,989	0.570	0.821	0.136810	0.974551	5.38
	10	0.447	1,737	5,587	1,695	28,051	0.506	0.834	0.237165	0.943018	4.16
	25	0.321	3,544	3,780	5,145	24,601	0.408	0.867	0.483889	0.827036	2.80
	50	0.193	5,586	1,738	12,100	17,646	0.316	0.910	0.762698	0.593223	1.87
	75	0.086	6,651	673	20,172	9,574	0.248	0.934	0.908110	0.321858	1.34
	90	0.060	7,099	225	25,234	4,512	0.220	0.953	0.969279	0.151684	1.14

*Compared to pre-test probability of around 18% based on prevalence.

Note: TP: true positives; FP: false positives; FN: false negatives; TN: true negatives; Se: sensitivity; Sp: specificity; LR+: positive likelihood ratio; NPV: negative predictive value; PPV: positive predictive value (post-test probability)

Table 5.2. Simulation metrics for predictions classified daily using XGBoost.

Top BZRA dispenses	Threshold	TP	FN	FP	TN	PPV	NPV	Se	Sp	LR+
10	0.512	6	123	5	667	0.517	0.844	0.044202	0.992107	5.60
20	0.433	10	104	11	630	0.465	0.858	0.085457	0.982472	4.88
50	0.297	18	68	33	519	0.351	0.884	0.207843	0.94002	3.47
100	0.168	25	36	72	350	0.247	0.907	0.406822	0.829704	2.39

Note: predictions were classified daily for 365 days, then mean values of TP, FP, TN and FNs were used for subsequent calculations

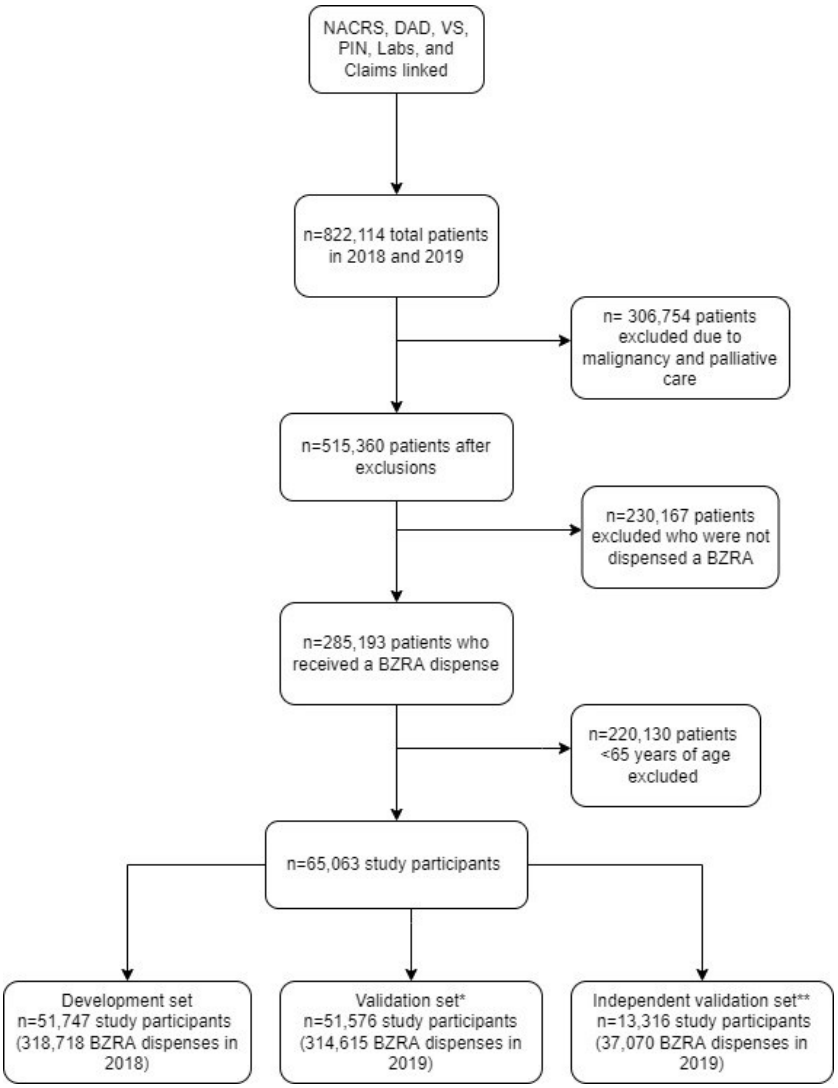
BZRA: benzodiazepine receptor modulator; TP: true positives; FP: false positives; FN: false negatives; TN: true negatives; Se: sensitivity; Sp: specificity; LR+: positive likelihood ratio; NPV: negative predictive value; PPV: positive predictive value (post-test probability)

Table 5.3. Simulation metrics for predictions classified quarterly using XGBoost.

	Top BZRA Dispenses	Threshold	TP	FN	FP	TN	PPV	NPV	Se	Sp	LR+
Quarter 1	500	0.631	344	5,669	157	26,818	0.687	0.825	0.057	0.994	9.83
	1,000	0.559	629	5,384	372	26,603	0.628	0.832	0.105	0.986	7.59
	2,000	0.481	1,125	4,888	876	26,099	0.562	0.842	0.187	0.968	5.76
	3,000	0.432	1,526	4,487	1,476	25,499	0.508	0.850	0.254	0.945	4.64
	5,000	0.367	2,286	3,727	2,716	24,259	0.457	0.867	0.380	0.899	3.78
	10,000	0.271	3,706	2,307	6,300	20,675	0.370	0.900	0.616	0.766	2.64
Quarter 2	500	0.588	316	5,439	185	27,179	0.631	0.833	0.055	0.993	8.12
	1,000	0.518	549	5,022	452	26,721	0.548	0.842	0.099	0.983	5.92
	2,000	0.438	963	4,307	1,038	25,664	0.481	0.856	0.183	0.961	4.70
	3,000	0.384	1,295	3,710	1,708	24,466	0.431	0.868	0.259	0.935	3.97
	5,000	0.309	1,831	2,682	3,171	21,939	0.366	0.891	0.406	0.874	3.21
	10,000	0.166	2,552	1,002	7,452	14,779	0.255	0.937	0.718	0.665	2.14
Quarter 3	500	0.585	293	5,013	208	27,518	0.585	0.846	0.055	0.992	7.36
	1,000	0.505	525	4,498	477	26,826	0.524	0.856	0.105	0.983	5.98
	2,000	0.415	897	3,631	1,105	25,217	0.448	0.874	0.198	0.958	4.72
	3,000	0.358	1,173	2,943	1,831	23,405	0.390	0.888	0.285	0.927	3.93
	5,000	0.276	1,564	1,839	3,443	19,613	0.312	0.914	0.460	0.851	3.08
	10,000	0.097	1,772	404	8,241	8,878	0.177	0.956	0.814	0.519	1.69
Quarter 4	500	0.582	224	3,903	278	28,946	0.446	0.881	0.054	0.990	5.71
	1,000	0.498	415	3,450	587	27,872	0.414	0.890	0.107	0.979	5.21
	2,000	0.405	697	2,682	1,305	25,517	0.348	0.905	0.206	0.951	4.24
	3,000	0.341	902	2,079	2,100	23,015	0.300	0.917	0.303	0.916	3.62
	5,000	0.249	1,162	1,137	3,841	17,828	0.232	0.940	0.505	0.823	2.85
	10,000	0.069	1,132	116	8,873	3,701	0.113	0.970	0.907	0.294	1.29

BZRA: benzodiazepine receptor modulator; TP: true positives; FP: false positives; FN: false negatives; TN: true negatives; Se: sensitivity; Sp: specificity; LR+: positive likelihood ratio; NPV: negative predictive value; PPV: positive predictive value (post-test probability)

Figure 5.1. Study participant flow diagram used for developing and validating machine learning models.



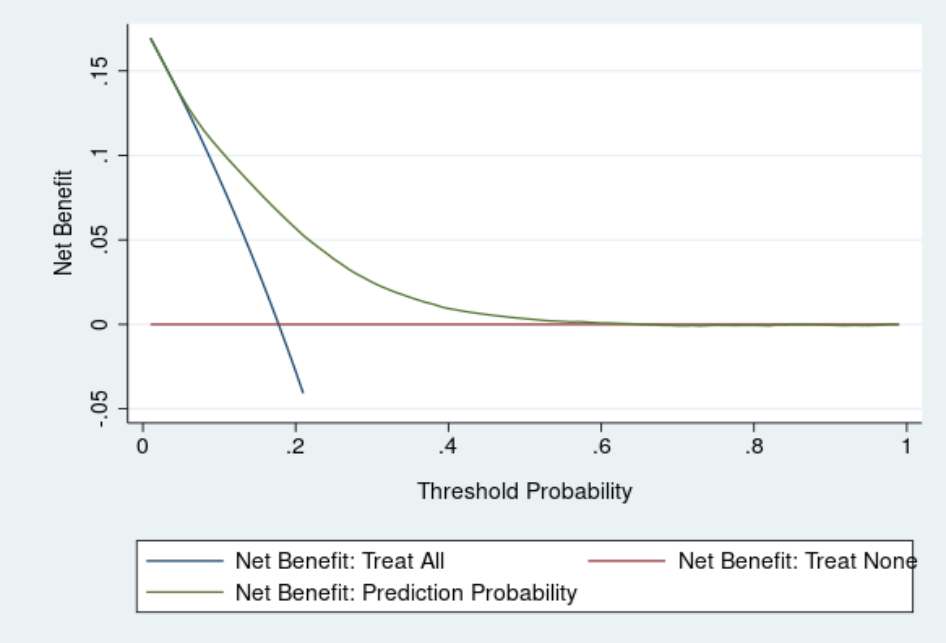
Note: NACRS: National Ambulatory Care Reporting System; DAD: Discharge Abstract Database; VS: Vital Statistics; PIN: Pharmaceutical Information Network; Claims: Physician Claims; Labs: Provincial Laboratory database; BZRA: benzodiazepine receptor modulator

*Participants in this set could be represented in both the development and validation set. The ML model will classify both new and previous patients.

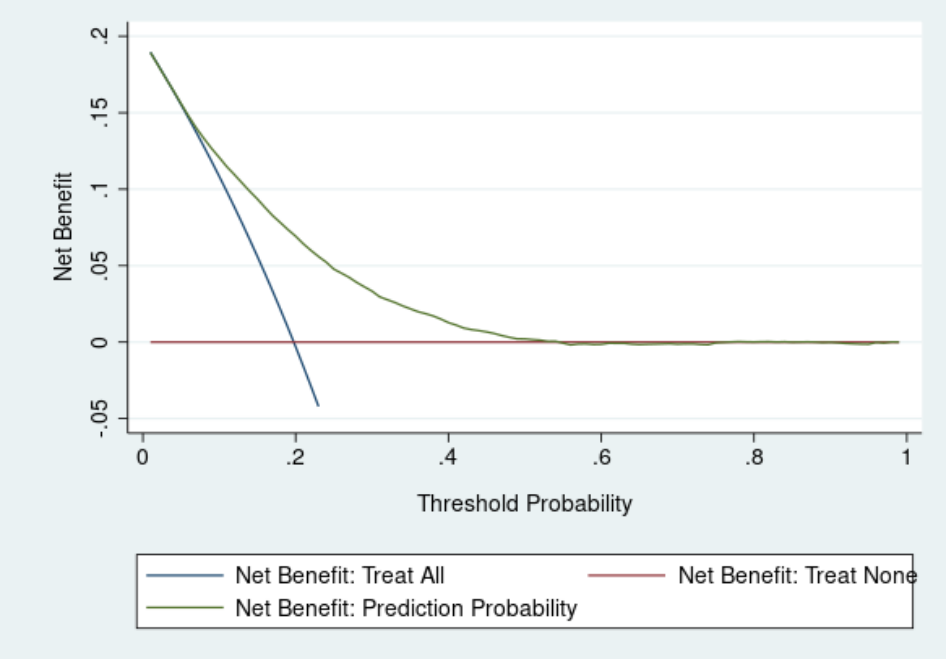
**Participants in this validation set are not included in the development set. This validation set represents out of sample independent BZRA patients in 2019. The ML model will classify only new patients.

Figure 5.2. Decision curve/net benefit analysis for all-cause outcomes and the XGBoost classifier for the entire validation set (A), and out of sample validation set (B).

(A)



(B)



Note: "Prediction Probability" in the legend refers to XGBoost classifier predictions. "Treat all" refers to if all dispenses were intervened on and "treat none" refers to if none of the dispenses were intervened on.

Figure 5.3. Simulation of events per 100 daily benzodiazepine receptor modulator dispenses using the XGBoost classifier stratified by percentile of predicted risk. Baseline risk corresponds to the pre-test probability.

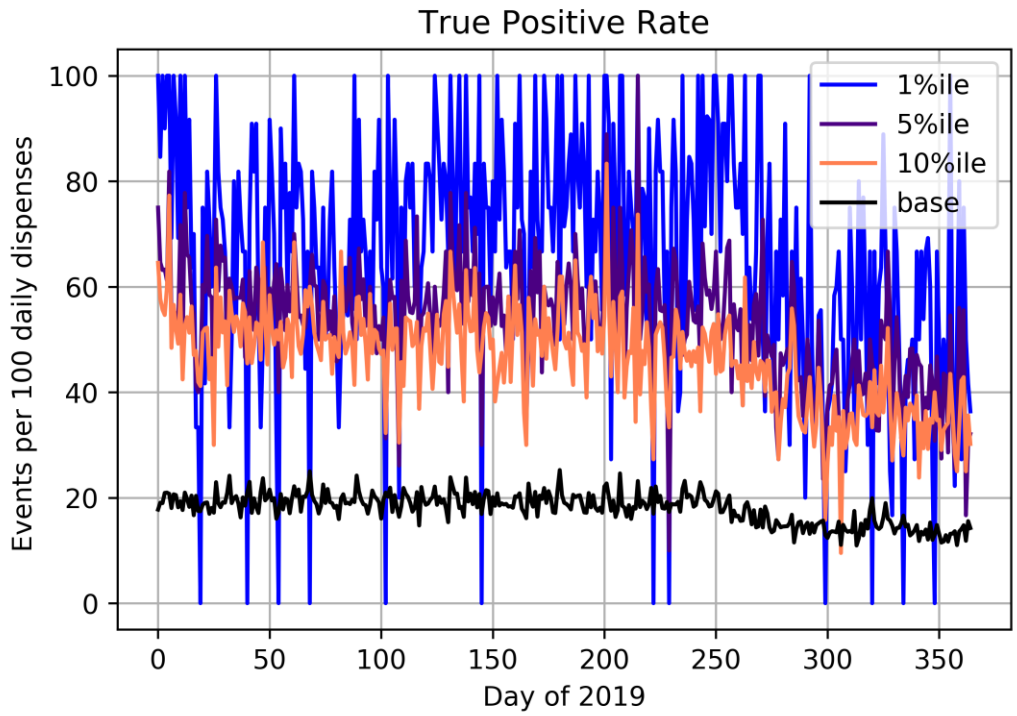
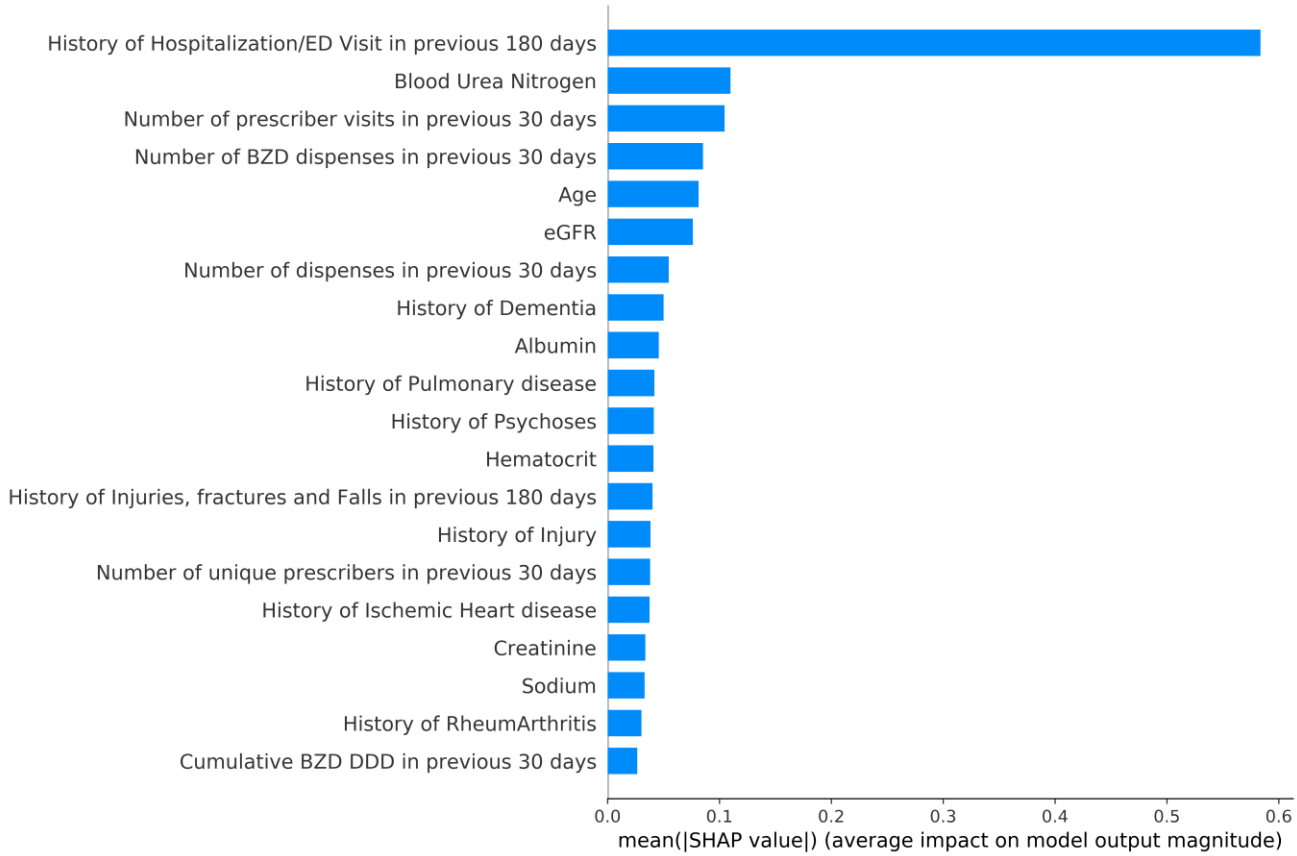


Figure 5.4. Variable importance graph of the XGBoost classifier predicting all-cause outcomes.



Appendix to Chapter 5

Table 5.4. Anatomical Therapeutic Chemical classification of BZRA molecules used for this study and candidate predictor categories used to develop the ML model.

Category (data source)	Description
ATC codes used to identify BZRAs (PIN)	Level 4 ATC code N05BA, N05CD, N05CF, N03AE01 (clonazepam)
BZRA molecules used in this study (PIN)	zopiclone, zolpidem, eszopiclone, zaleplon, alprazolam, bromazepam, chlordiazepoxide, clobazam, clonazepam, clorazepate, diazepam, flurazepam, lorazepam, midazolam, nitrazepam, oxazepam, temazepam, triazolam
Demographic information (PIN)	age, sex
Drug utilization history (PIN)	drug dispenses in past 30 days using level 3 ATC codes, oral morphine equivalents, concurrent use of opioids and BZRAs defined as at least 7 days of cumulative concurrent use in the 30 days prior to BZRA dispensation, number of dispensations and unique molecules of BZRAs, BZRA DDDs
Healthcare utilization (PIN)	number of prescribers and pharmacies, prior inpatient admissions
ICD based co-morbidities (DAD, NACRS, Claims)	co-morbidity flags based on the past 5 years of physician claims, hospitalizations, and emergency visits: alcohol use disorder, cardiac arrhythmias, blood loss anemia, congestive heart failure, cancer, lymphoma, coagulopathy, deficiency anemia, depression, diabetes, drug abuse, fluid disorder, HIV/aids, hypertension, hypothyroidism, liver disorder, other neurological disorders, obesity, paralysis, psychoses, peptic ulcer disease, pulmonary circulation disorders, chronic pulmonary disease, peripheral vascular disorders, renal failure, rheumatoid arthritis, valvular disease, weight loss, stroke, ischemic heart disease, hyperlipidemia, tobacco use, Alzheimer's disease, delirium, difficulty walking, falls, incontinence, abnormality of gait, senility without mention of psychosis, vascular dementia, pressure ulcer, dementia, malaise
Lab results (Provincial Laboratory)	Most recent result within 1 year of BZRA dispensation. Includes hemoglobin A1C, albumin, ALT, blood urea nitrogen, creatinine, eGFR, HDL, hematocrit, LDL, electrolytes, cholesterol ratio, total cholesterol, triglycerides

Note: BZRA: benzodiazepine receptor modulator; ATC- Anatomical Therapeutic Chemical classification (https://www.whocc.no/atc_ddd_index); PIN- Pharmaceutical Information Network; DAD: Discharge Abstract Database; NACRS: National Ambulatory Care Reporting System; Claims: Physician Claims; ICD: International Statistical Classification of Diseases and Related Health Problems; DDD: defined daily dose

Table 5.5. ICD-10 codes used to define our cause specific composite outcome.

Composite Outcome	ICD-10*
Poisonings	X40-49; Y10-19; X60-84; T36-T65
Injuries	S00-99; T00-T35; T67; T68; T71; T73; T74
Falls	W00-W19
Accidents	W20-99; V09-V19; V20.X-V28.X (X=.0 OR .4); V30.X-V70.X (X=.0, .5); V80; V83.X-V86.X (X=.0, .5); V87-V94;

* International Statistical Classification of Diseases and Related Health Problems

Table 5.6. Co-morbidity characteristics of those who did and did not experience an all-cause outcome using 2018-2019 data.

Characteristic	Number without event (n=60,648)	Percent	Number with event (n=30,663)	Percent
Mean Age (SD)	77.6 (9.2)	(--)	76.6 (8.4)	(--)
Female	40,217	66.3	20,065	65.4
Male	20,431	33.7	10,598	34.6
Alcohol use disorder	10,339	17.0	6,139	20.0
Arrhythmia	28,171	46.5	16,069	52.4
Anemia	361	0.6	298	1.0
Heart Failure	9,819	16.2	7,165	23.4
Coagulopathy	2,794	4.6	2,014	6.6
Iron deficiency anemia	8,850	14.6	5,634	18.4
Depression	33,693	55.6	18,491	60.3
Diabetes	18,349	30.3	10,569	34.5
History of drug abuse	10,751	17.7	6,229	20.3
Fluid and electrolyte disorder	13,560	22.4	9,258	30.2
Hypertension	46,129	76.1	24,827	81.0
Hypothyroidism	15,838	26.1	8,444	27.5
Injury	46,920	77.4	25,681	83.8
Liver disorder	3,533	5.8	2,106	6.9
Lymphoma	727	1.2	430	1.4
Neurologic disorder	42,586	70.2	23,093	75.3
Obesity	11,037	18.2	6,196	20.2
History of poisoning	4,103	6.8	2,730	8.9
Psychoses	8,532	14.1	5,567	18.2
Renal disorder	8,076	13.3	5,376	17.5
Rheumatoid disorder	22,372	36.9	13,241	43.2
Cancer	5,456	9.0	3,148	10.3
HIV	59	0.1	50	0.2
Paralysis	1,326	2.2	971	3.2
Peptic ulcer disease	2,840	4.7	1,798	5.9
Pulmonary circulation disorder	3,463	5.7	2,614	8.5
Pulmonary disease	21,557	35.5	13,155	42.9
Peripheral vascular disorder	6,460	10.7	4,436	14.5
Valvular disease	3,625	6.0	2,583	8.4
Weight loss	6,359	10.5	4,131	13.5

Note: p-value<0.001 for all comparisons except for male and female (p=0.008); this analysis was done at the participant level. Participants could be represented in either or both groups; (--): not estimable

Table 5.7. Co-morbidity characteristics of participants in the development and validation sets.

Characteristic	Development set 2018 (n=51,747)	Percent	Validation set 2019 (n=51,576)	Percent
Mean Age (SD)*	77.5 (9.0)	(--)	77.3 (9.0)	(--)
Male	17,307	33.4	17,390	33.7
Female	34,440	66.6	34,186	66.3
Alcohol use disorder*	8,470	16.4	9,053	17.6
Arrhythmia*	23,559	45.5	24,433	47.4
Blood loss anemia	320	0.6	327	0.6
Heart failure*	8,373	16.2	8,669	16.8
Coagulopathy	2,439	4.7	2,457	4.8
Iron deficiency anemia*	7,464	14.4	7,704	14.9
Depression*	28,057	54.2	29,303	56.8
Diabetes*	15,320	29.6	15,884	30.8
History of drug abuse*	8,501	16.4	9,580	18.6
Fluid and electrolyte disorder*	11,329	21.9	11,922	23.1
Hypertension*	39,372	76.1	39,614	76.8
Hypothyroidism*	13,336	25.8	13,664	26.5
History of injury*	39,269	75.9	40,574	78.7
Liver disorder*	2,780	5.4	3,147	6.1
Lymphoma	642	1.2	647	1.3
Neurologic disorder*	35,467	68.5	36,661	71.1
Obesity*	9,245	17.9	9,607	18.6
History of poisoning*	3,408	6.6	3,635	7.0
***Psychoses**	6,982	13.5	7,423	14.4
Renal disorder*	6,640	12.8	7,265	14.1
Rheumatoid disorder*	18,581	35.9	19,691	38.2
Cancer*	4,866	9.4	4,595	8.9
HIV	51	0.1	54	0.1
History of paralysis	1,083	2.1	1,165	2.3
History of peptic ulcer disease	2,381	4.6	2,478	4.8
Pulmonary circulation disorder	2,975	5.7	3,084	6.0
Pulmonary disease*	18,164	35.1	18,804	36.5
Peripheral vascular disorder*	5,317	10.3	5,801	11.2
Valvular disorder	3,127	6.0	3,226	6.3
Weight loss*	5,200	10.0	5,485	10.6

Note: Study participants can be in either or both groups; (--): not estimable

*p-value <0.05

Table 5.8. Distribution of BZRA dispenses without a lab test among commonly ordered lab test results.

	Number of BZRA dispenses with missing labs					
Lab	**2018**	**%**	**2019**	**%**	**2018 and 2019**	**%**
Hemoglobin A1C	161,652	50.7	151,993	48.3	313,645	49.5
Albumin	203,602	63.9	201,110	63.9	404,712	63.9
ALT	123,491	38.7	122,880	39.1	246,371	38.9
BUN	183,844	57.7	183,377	58.3	367,221	58.0
Creatinine	50,423	15.8	51,630	16.4	102,053	16.1
eGFR	223,764	70.2	106,263	33.8	330,027	52.1
HDL	190,262	59.7	185,253	58.9	375,515	59.3
Hematocrit	59,659	18.7	60,960	19.4	120,619	19.0
LDL	192,834	60.5	187,090	59.5	379,924	60.0
Potassium	69,405	21.8	70,048	22.3	139,453	22.0
Sodium	71,174	22.3	71,505	22.7	142,679	22.5
Total Cholesterol	188,696	59.2	183,692	58.4	372,388	58.8
Triglycerides	189,861	59.6	184,914	58.8	374,775	59.2

Note: BZRA: benzodiazepine receptor modulator; total BZRA dispenses for 2018 and 2019 are 318,718 and 314,615, respectively.

Table 5.9. Area under the receiver operating characteristic curve (AUROC) for various ML algorithms using the entire 2019 validation set.

Algorithm	Development set (2018)	Validation set (2019)
Logistic Regression	0.667	0.672
Gradient Boosting Classifier	0.753	0.741
Linear Support Vector	0.674	0.675
Decision Tree Classifier	0.771	0.712
Random Forest Classifier	0.782	0.739
Ridge Classifier	0.752	0.741
Multi-layer Perceptron	0.74500	0.736
XGBoost	0.801	0.749
AdaBoost Classifier	0.748	0.740

Note: all-cause outcomes were assessed.

Table 5.10. C-statistics for the all-cause outcome XGBoost classifier using fewer datasets and for the composite outcome and its individual components.

Category	Development set (2018)	Validation Set (2019 full set)	Number of Features
All features	0.80	0.75**	326
Demographic and drug histories only	0.71	0.66	252
All features excluding lab test results	0.79	0.75	312
Composite outcome*	0.76	0.69	326
Accidents*	0.95	0.62	326
Poisonings*	0.93	0.86	326
Falls*	0.78	0.70	326
Injuries*	0.91	0.68	326

*All datasets were used to train the XGBoost model for these outcomes.

**The corresponding c-statistic for the out of sample validation set (new patients) was 0.74.

Table 5.11. All-cause outcome prediction metrics stratified by top highest risk dispenses for the XGBoost classifier measured at the end of 2019 on both validation sets.

	Top BZRA dispenses	Threshold	TP	FN	FP	TN	PPV*	NPV	Se	Sp	LR+
Entire validation set	**10**	0.984	10	55,918	1	258,686	0.909	0.822	0.000179	0.999996	46.25
	20	0.979	21	55,907	2	258,685	0.913	0.822	0.000375	0.999992	48.57
	50	0.973	49	55,879	5	258,682	0.907	0.822	0.000876	0.999981	45.33
	100	0.965	92	55,836	12	258,675	0.885	0.822	0.001645	0.999954	35.46
	500	0.916	467	55,461	50	258,637	0.903	0.823	0.008350	0.999807	43.20
	1000	0.843	936	54,992	212	258,475	0.815	0.825	0.016736	0.999180	20.42
	5000	0.608	3,692	52,236	2,043	256,644	0.644	0.831	0.066013	0.992102	8.36
	10000	0.528	6,436	49,492	4,667	254,020	0.580	0.837	0.115077	0.981959	6.38
Independent validation set	**10**	0.965	11	7,313	0	29,746	1.000	0.803	0.001502	1.000000	(--)
	20	0.957	21	7,303	0	29,746	1.000	0.803	0.002867	1.000000	(--)
	50	0.928	47	7,277	4	29,742	0.922	0.803	0.006417	0.999866	47.72
	100	0.852	91	7,233	11	29,735	0.892	0.804	0.012425	0.999630	33.60
	500	0.668	366	6,958	208	29,538	0.638	0.809	0.049973	0.993007	7.15
	1000	0.589	659	6,665	442	29,304	0.599	0.815	0.089978	0.985141	6.06
	5000	0.390	2,433	4,891	2,864	26,882	0.459	0.846	0.332196	0.903718	3.45
	10000	0.290	4,082	3,242	6,630	23,116	0.381	0.877	0.557346	0.777113	2.50

*Compared to pre-test probability of around 18% based on prevalence.

Note: BZRA: benzodiazepine receptor modulator; TP: true positives; FP: false positives; FN: false negatives; TN: true negatives; Se: sensitivity; Sp: specificity; LR+: positive likelihood ratio; NPV: negative predictive value; PPV: positive predictive value (post-test probability); (--): not estimable

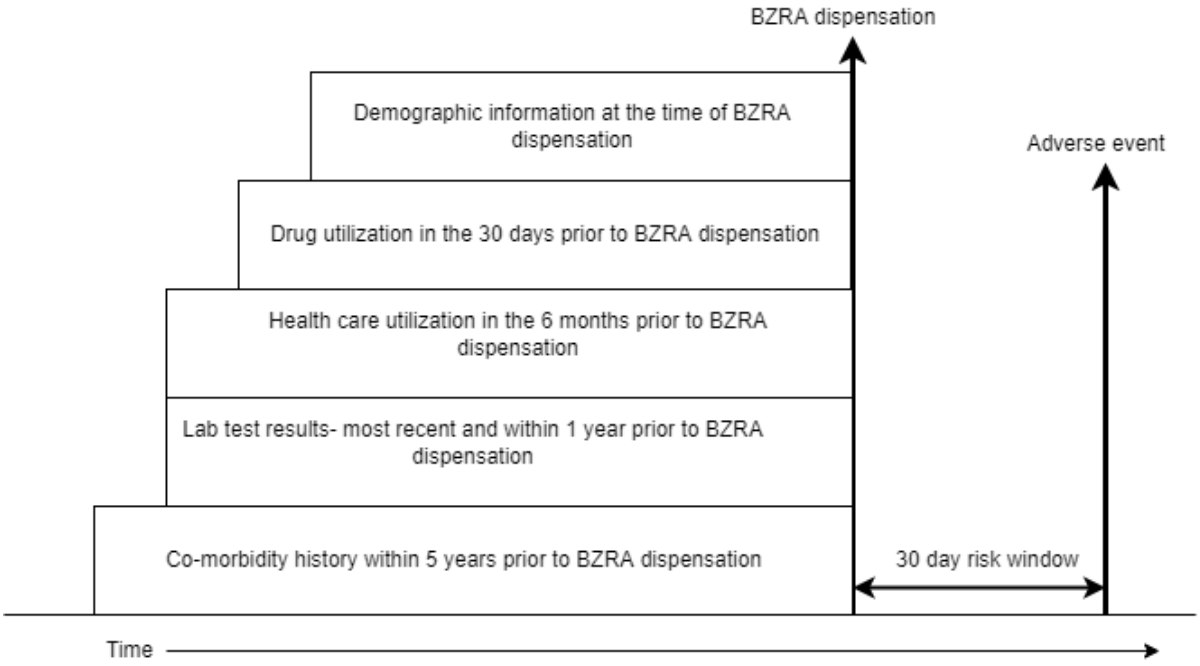
Table 5.12. All-cause outcome prediction metrics stratified by absolute thresholds for the XGBoost classifier measured at the end of 2019 on both validation sets.

	Threshold	TP	FN	FP	TN	PPV*	NPV	Se	Sp	LR+
Entire validation set	0	55,928	0	258,687	0	0.178	(--)	1	0	1.00
	0.1	49,328	6,600	149,965	108,722	0.248	0.943	0.881991	0.420284	1.52
	0.2	41,401	14,527	93,954	164,733	0.306	0.919	0.740255	0.636804	2.04
	0.3	28,227	27,701	46,521	212,166	0.378	0.885	0.504702	0.820165	2.81
	0.4	15,521	40,407	17,943	240,744	0.464	0.856	0.277518	0.930638	4.00
	0.5	7,856	48,072	6,242	252,445	0.557	0.840	0.140466	0.97587	5.82
	0.6	3,876	52,052	2,226	256,461	0.635	0.831	0.069303	0.991395	8.05
	0.7	2,058	53,870	920	257,767	0.691	0.827	0.036797	0.996444	10.35
	0.8	1,201	54,727	351	258,336	0.774	0.825	0.021474	0.998643	15.83
	0.9	560	55,368	77	258,610	0.879	0.824	0.010013	0.999702	33.64
	1	0	55,928	0	258,687	(--)	0.822	0	1	(--)
Independent validation set	0	7,324	0	29,746	0	0.198	(--)	1	0	1.00
	0.1	6,472	852	18,273	11,473	0.262	0.931	0.88367	0.385699	1.44
	0.2	5,515	1,809	11,764	17,982	0.319	0.909	0.753004	0.604518	1.90
	0.3	3,903	3,421	6,125	23,621	0.389	0.873	0.532906	0.79409	2.59
	0.4	2,320	5,004	2,612	27,134	0.470	0.844	0.316767	0.91219	3.61
	0.5	1,245	6,079	1,041	28,705	0.545	0.825	0.169989	0.965004	4.86
	0.6	612	6,712	401	29,345	0.604	0.814	0.083561	0.986519	6.20
	0.7	270	7,054	161	29,585	0.626	0.807	0.036865	0.994588	6.81
	0.8	133	7,191	27	29,719	0.831	0.805	0.018159	0.999092	20.01
	0.9	65	7,259	7	29,739	0.903	0.804	0.008875	0.999765	37.71
	1	0	7,324	0	29,746	(--)	0.802	0	1	(--)

*Compared to pre-test probability of around 18% based on prevalence.

Note: TP: true positives; FP: false positives; FN: false negatives; TN: true negatives; Se: sensitivity; Sp: specificity; LR+: positive likelihood ratio; NPV: negative predictive value; PPV: positive predictive value (post-test probability); (--): not estimable

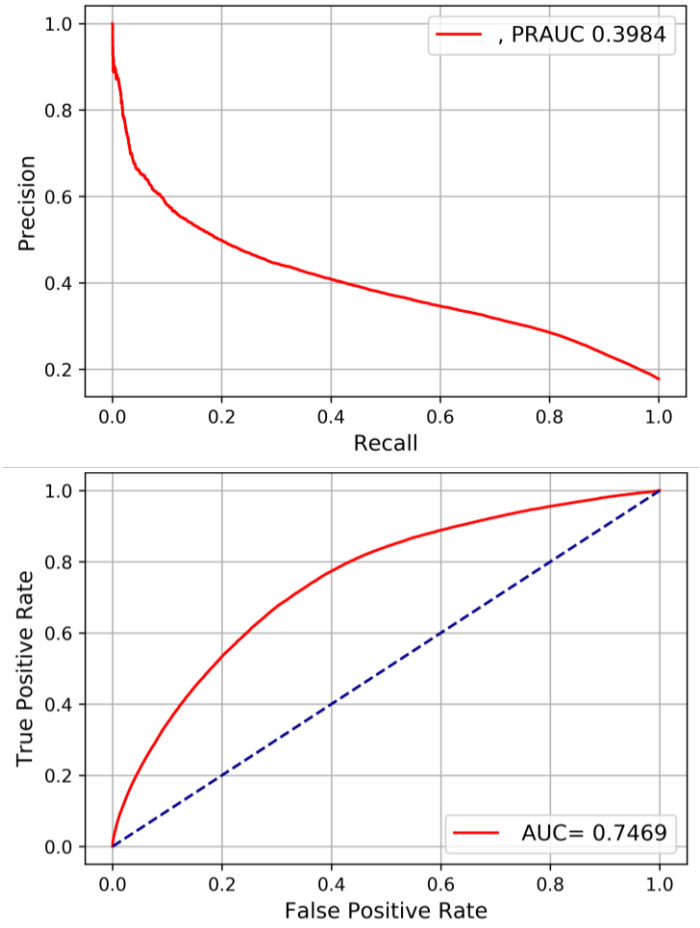
Figure 5.5. Feature generation timeline



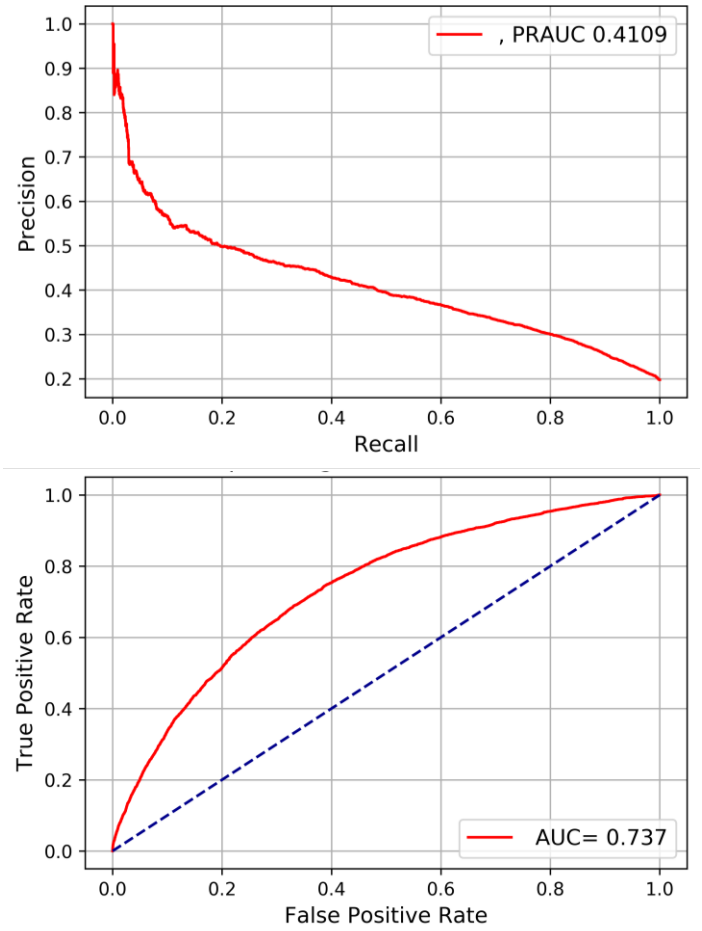
BZRA: benzodiazepine receptor modulator; adverse event: hospitalization, emergency department visit, or death

Figure 5.6. Precision-recall and area under the receiver operating characteristic curves for the XGBoost classifier for the entire validation set (A) and the out of sample validation set (B).

(A)



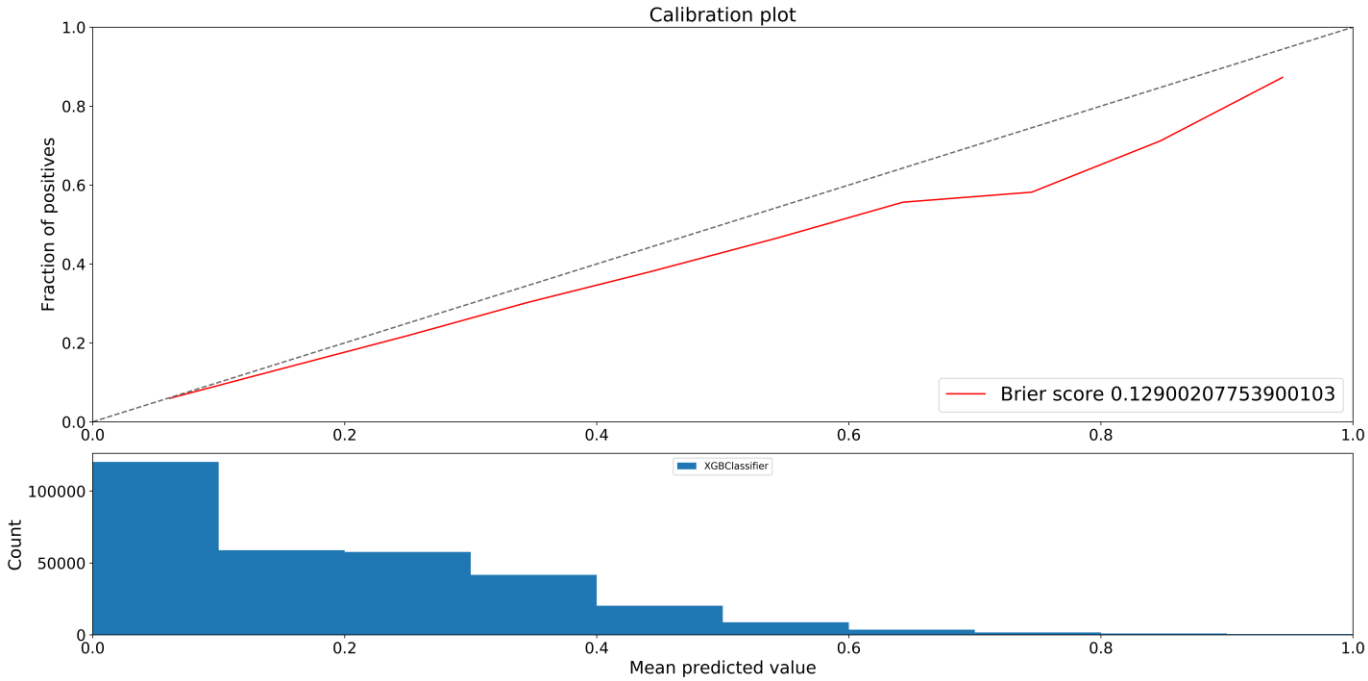
(B)



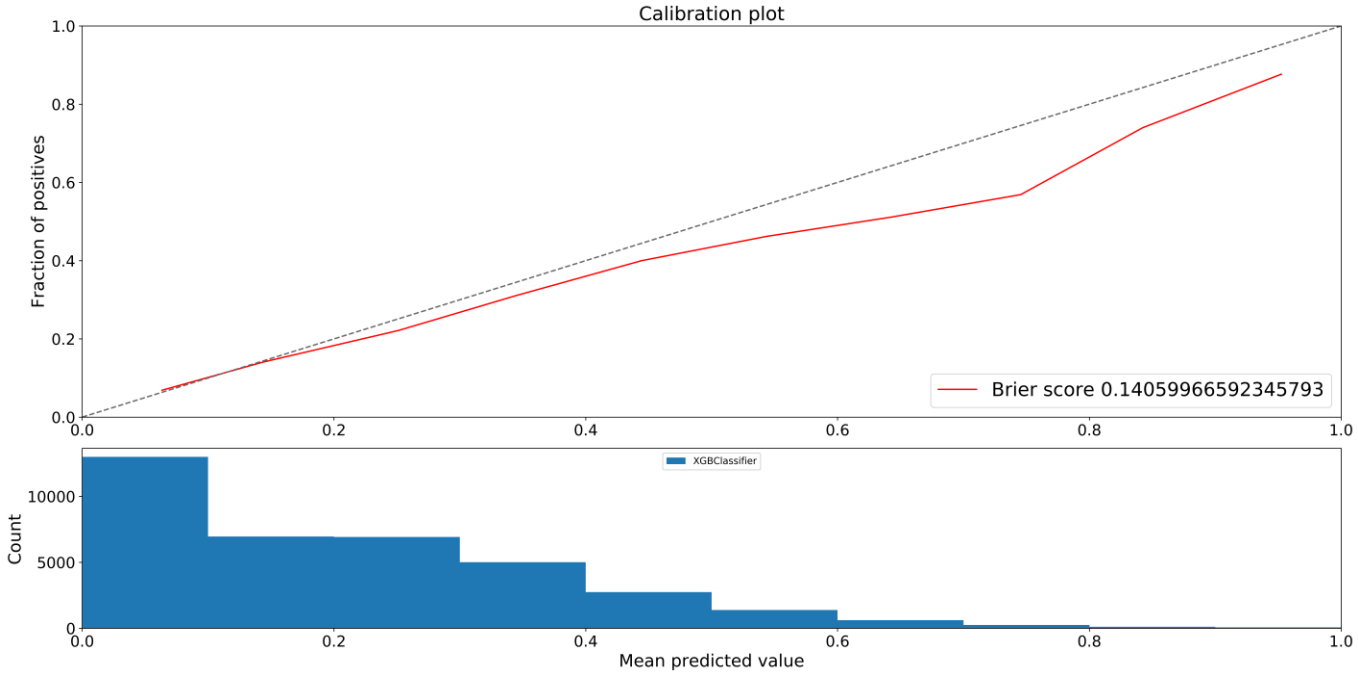
Note: These curves are for all-cause outcomes.

Figure 5.7. XGBoost calibration plots for all-cause outcomes using the entire validation set (A) and the out of sample validation set (B).

(A)



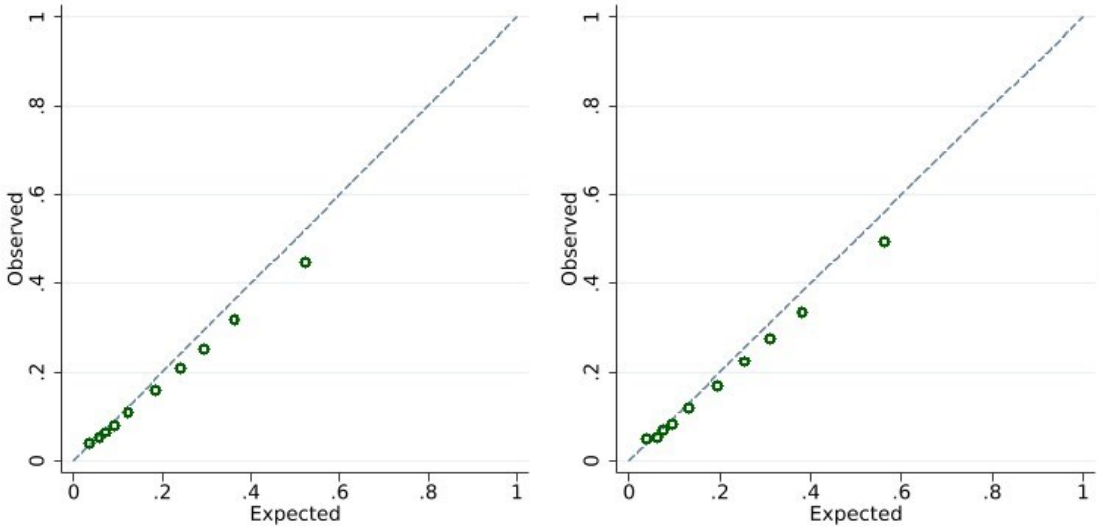
(B)



Note: Most of the predictions were classified as low risk in this data set.

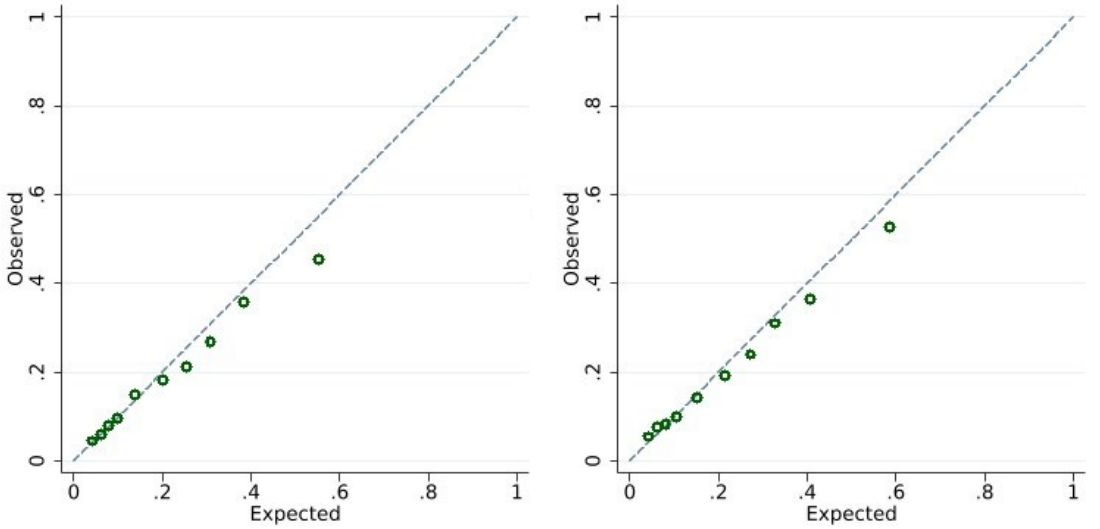
Figure 5.8. All-cause outcome calibration for XGBoost across sex for entire validation set (A) and out of sample validation set (B).

(A)



Males **Females**

(B)



Males **Females**

Figure 5.9. Negative predicted value vs predicted probability thresholds for the XGBoost classifier predicting all-cause outcomes.

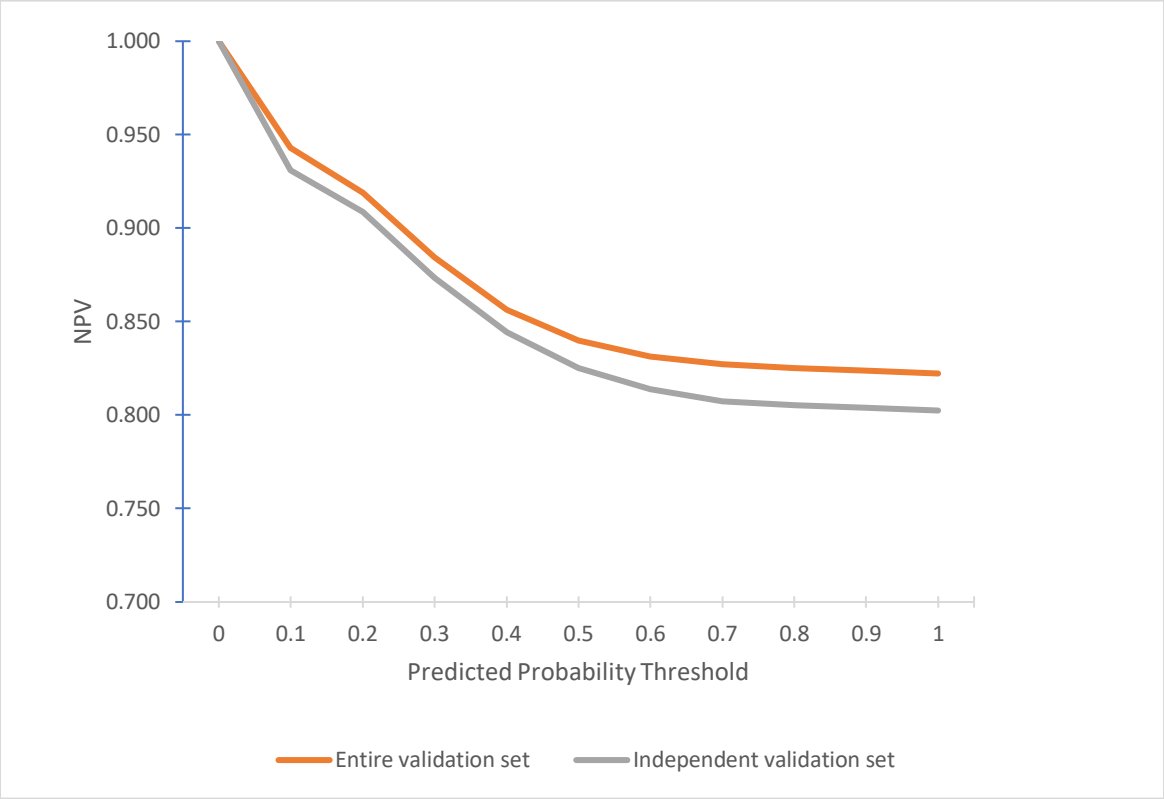
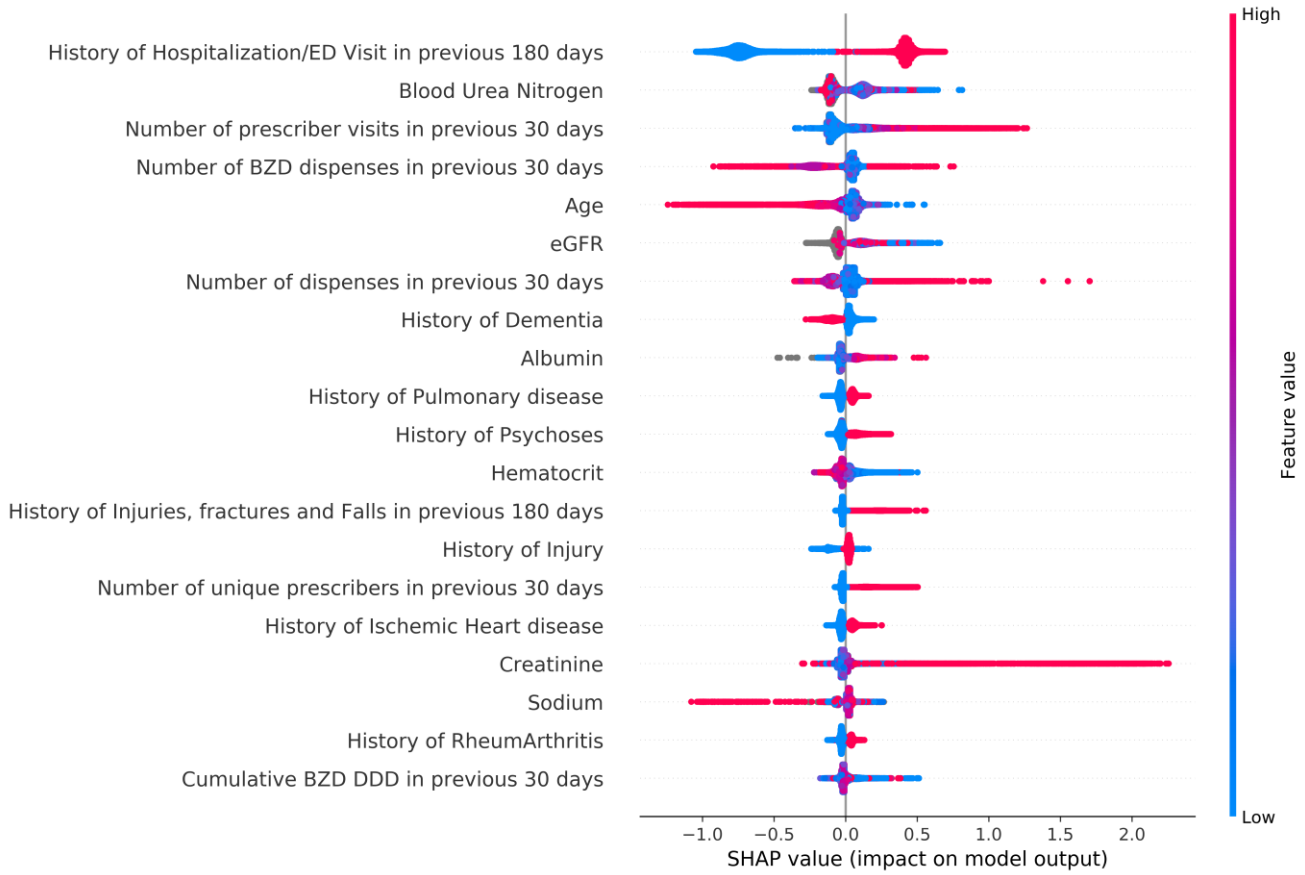


Figure 5.10. SHAP plot indicating influence of various predictors on risk of all-cause outcomes*.



*Red indicates higher values of the predictors while blue indicates lower values and lines to the right of the 0.0 on the x-axis are associated with our outcome while those to the left are less associated with our outcome. For example, a history of admission in the previous 180 days (red colour and to the right of the 0.0 on the x-axis) is predictive of outcomes to varying extents.

SHAP: Shapley Additive Explanations; BZD: benzodiazepine receptor modulator; DDD: defined daily dose

References for Chapter 5

11. 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.
12. Bastanlar Y, Ozuysal M. Introduction to machine learning. *Methods in molecular biology (Clifton, NJ).* 2014;1107:105-128.
13. 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.
14. Alberta Machine Intelligence Institute. Machine Learning Process Lifecycle. In:2019.
15. Shah NH, Milstein A, Bagley P, Steven C. Making Machine Learning Models Clinically Useful. *JAMA.* 2019;322(14):1351-1352.
16. Morgenstern JD, Buajitti E, O'Neill M, et al. Predicting population health with machine learning: a scoping review. *BMJ Open.* 2020;10(10):e037860.
17. 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.
18. Luo W, Phung D, Tran T, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res.* 2016;18(12):e323.
19. 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.
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. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *Bmj.* 2016;352:i6.
26. Cunningham CM, Hanley GE, Morgan S. Patterns in the use of benzodiazepines in British Columbia: examining the impact of increasing research and guideline cautions against long-term use. *Health Policy.* 2010;97(2):122-129.
28. ChooseWiselyCanada. The Canadian Geriatrics Society has developed a list of 5 things physicians and patients should question in geriatrics [Internet]. https://choosingwiselycanada.org/geriatrics/.

29. Ngiam KY, Khor W. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology.* 2019;20(5):e262-e273.

32. Cognilytica. Cognitive Project Management for Artifical Intelligence Methodology. In:2020.

34. Mooney SJ, Pejaver V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annual Review of Public Health.* 2018;39(1):95-112.

37. 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.

38. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European heart journal.* 2014;35(29):1925-1931.

42. Frizzell JD, Liang L, Schulte PJ, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA cardiology.* 2017;2(2):204-209.

45. Donders ART, Van Der Heijden GJ, Stijnen T, Moons KG. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology.* 2006;59(10):1087-1091.

46. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. *Journal of Clinical Epidemiology.*

50. 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.

51. Hu Z, Melton GB, Arsoniadis EG, Wang Y, Kwaan MR, Simon GJ. Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *Journal of Biomedical Informatics.* 2017;68:112-120.

52. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological methods.* 2002;7(2):147.

53. Sharafoddini A, Dubin JA, Maslove DM, Lee J. A new insight into missing data in intensive care unit patient profiles: Observational study. *JMIR medical informatics.* 2019;7(1):e11605.

59. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA.* 2018;320(21):2199-2200.

66. 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.

68. 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.

77. World health Organization. Classification of Diseases (ICD). 2019; https://www.who.int/classifications/icd/icdonlineversions/en/. Accessed Jun 2020.

79. 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.

80. World Health Organization. International language for drug utilization research, ATC/DDD. 2020; https://www.whocc.no/. Accessed Jun 2020, 2020.

82. 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.

87. 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.

88. 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.

89. 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.

91. Molnar C. *Interpretable machine learning. A Guide for Making Black Box Models Explainable.* 2019.

92. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Paper presented at: Advances in neural information processing systems2017.

94. 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.

95. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Paper presented at: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining2016.

96. The pandas development team. pandas-dev/pandas: Pandas. 2020; https://doi.org/10.5281/zenodo.3509134, Jan 2021.

107. Shin S, Austin PC, Ross HJ, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC Heart Failure.* 2021;8(1):106-115.

109. Alberta College of Pharmacy. 2019; https://abpharmacy.ca/. Accessed Sept 2019.

110. Canadian Institute for Health Information. 2019; https://www.cihi.ca/en.

113. Jiang W, Siddiqui S, Barnes S, et al. Readmission Risk Trajectories for Patients With Heart Failure Using a Dynamic Prediction Approach: Retrospective Study. *JMIR medical informatics.* 2019;7(4):e14756-e14756.

118. Retrum JH, Boggs J, Hersh A, et al. Patient-identified factors related to heart failure readmissions. *Circ Cardiovasc Qual Outcomes.* 2013;6(2):171-177.

119. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research.* 2011;12:2825-2830.

120. XGBoost. Python API Reference. https://xgboost.readthedocs.io/en/latest/python/python_api.html#module-xgboost.sklearn. Accessed August 2021.

128. van Smeden M, Groenwold RHH, Moons KGM. A cautionary note on the use of the missing indicator method for handling missing data in prediction research. *Journal of Clinical Epidemiology.*

130. Mitnitski A, Howlett SE, Rockwood K. Heterogeneity of Human Aging and Its Assessment. *The Journals of Gerontology: Series A.* 2016;72(7):877-884.

131. Liu L-F. The Health Heterogeneity of and Health Care Utilization by the Elderly in Taiwan. *International Journal of Environmental Research and Public Health.* 2014;11(2):1384-1397.

134. Alberta Health. Alberta Health Services performance review : summary report. 2020; https://open.alberta.ca/publications/alberta-health-services-performance-review-summary-report#summary.

135. Katzman MA, Bleau P, Blier P, et al. Canadian clinical practice guidelines for the management of anxiety, posttraumatic stress and obsessive-compulsive disorders. *BMC Psychiatry.* 2014;14 Suppl 1:S1.

136. Pottie K, Thompson W, Davies S, et al. Deprescribing benzodiazepine receptor agonists. 2018.

137. Canadian Pharmacists Association. RxTx. 2019; https://www.e-therapeutics.ca/search.

138. CPSA. Clinical Toolkit Benzodiazepines: Use and Taper. *CPSA.* 2015.

139. TOP TOP. Guideline for Adult Primary Insomnia [Internet]. 2010; http://www.topalbertadoctors.org/download/439/insomnia_management_guideline.pdf.

140. Weir DL, Samanani S, Gilani F, Jess E, Eurich DT. Benzodiazepine receptor agonist dispensations in Alberta: a population-based descriptive study. *CMAJ Open.* 2018;6(4):E678-E684.

141. College of Physicians and Surgeons of Alberta. Tracked Prescription Program. 2021; https://www.tppalberta.ca/.

142. American Geriatrics Society Beers Criteria Update Expert P. American Geriatrics Society updated Beers Criteria for potentially inappropriate medication use in older adults. *J Am Geriatr Soc.* 2012;60(4):616-631.

143. O'Mahony D. STOPP/START criteria for potentially inappropriate medications/potential prescribing omissions in older people: origin and progress. *Expert review of clinical pharmacology.* 2020;13(1):15-22.

144. Urquhart R, Giguere AM, Lawson B, et al. Rules to identify persons with frailty in administrative health databases. *Canadian Journal on Aging/La Revue canadienne du vieillissement.* 2017;36(4):514-521.

145. Jaeschke R, Guyatt GH, Sackett DL, Group EBMW. How to use an article about a diagnostic test. *Jama.* 1994;271(5):389-391.

Chapter 6: 30-day risk from prescribed opioids: creating a risk predictor for prescription drug monitoring programs using a machine learning approach

Importance: Machine-learning approaches can assist opioid stewardship by identifying high-risk opioid prescribing for potential interventions.

Objective: To develop a machine-learning model for deployment that can estimate risk of adverse outcomes within 30-days of an opioid dispensation as a potential component of prescription drug monitoring programs.

Design, Setting and Participants: This prognostic study used data from Alberta, Canada between 2018-2019. Participants included all patients over 18 who received at least one opioid dispensation from a community pharmacy within the province.

Exposure: Each opioid dispensation served as the unit of analysis.

Main Outcomes/Measures: We identified opioid-related adverse outcomes from linked administrative datasets. An XGBoost model was developed on 2018 data to predict risk of emergency department visit, hospitalization, or mortality within 30-days of an opioid dispensation; validation on 2019 data was done to evaluate model performance. We reported model discrimination, calibration, and other relevant metrics using daily and weekly predictions on both ranked predictions and predicted probability thresholds. A cost analysis describing potential savings based on our predictions was included.

Results: Total number of participants was 853,324 representing 6.1 million opioid dispensations with 145,016 events reported (2.3%). 77,326 events (2.6% pre-test probability) occurred within 30 days of a dispense in the validation set (XGBoost C-statistic 0.82). The top 0.1 percentile of predicted risk had a positive likelihood ratio (LR+) of 28.7 which translated to a post-test probability of 43.1%. In our simulations, the weekly measured predictions had higher LR+'s in both the highest risk dispenses and percentiles of predicted risk as compared to predictions measured daily. Net benefit analysis showed that using ML prediction was not

informative across the entire range of probability thresholds. Intervening on the highest ranked predictions demonstrated cost savings.

Conclusion:

Prescription drug monitoring programs can use machine-learning classifiers to identify patients at risk of adverse outcomes from opioids and potentially reduce health-care costs by intervening on high-ranked predictions. Better access to available administrative and clinical data could improve the prediction performance of ML classifiers, especially if probability thresholds are important, and thus expand opioid stewardship efforts and further reduce costs.

Introduction

Canada experiences some of the highest rates of opioid prescribing in the world, making prescription opioid use a key driver of the opioid crisis[7]. The consequences of prescribed opioids are well characterized[6,7,24,146]. As part of the response to the opioid crisis, health jurisdictions and regulators have implemented prescription drug monitoring programs (PDMPs) such as the Tracked Prescription Program (TPP) Alberta administered by the College of Physicians and Surgeons of Alberta (CPSA)[141], Canada. The CPSA, like many other health regulators, has a mandate to protect Albertans by guiding the medical profession and plays a major role in prescription opioid stewardship.

With the increase in digital health, there is a strong movement to integrate emerging digital technologies (e.g., machine learning) with medicine[11,71,147]. Health regulators like the CPSA are leading this trend by using data to optimize patient safety[141]. The Government of Alberta maintains a comprehensive infrastructure of administrative health data and the CPSA has limited access to certain datasets, namely community prescription dispensation records.

Supervised machine learning (ML)[11] is an approach that uses computer algorithms and the large amounts of available administrative data to build clinical prediction models, all within a well-defined framework[14]. Supervised ML trains on labelled data to develop prediction models that are specific to different populations and there are numerous published studies describing its use in clinical settings[16,17,37]. However, reporting of ML prediction performance metrics is still inconsistent in the literature[17] although guidelines are in the works[18]. To date, we are not aware of any health jurisdictions in Canada which use ML approaches in their opioid stewardship programs.

Building on our previous work[148], the objective of this study was to develop and validate a proof-of-concept XGBoost[95,120] ML model for use by the CPSA that can estimate the risk of adverse outcomes within 30-days of an opioid dispensation in Alberta, Canada. The ML model will be trained only on prescription drug records to simulate a potential ML classifier deployment by the CPSA that is aligned with the type of data TPP Alberta has access to. We will evaluate the ML model using the same performance metrics[18-20,38] as in our previous work and

provide a general description of a customizable online dashboard. Using the CPSA as an example, our analysis will provide interested prescription regulators with analytic options to assess the value and implementation of the ML classifier based on workload capacity and potential cost savings. Although our study was conducted for the CPSA using data from Alberta, the ML process allows for others in different jurisdictions to deploy their own population and data specific ML risk classifiers.

Methods

Study Design, Setting and Participants

This prognostic study used a supervised ML approach which trained an XGBoost model on opioid dispensations in Alberta, Canada between Jan 2018 – Dec 2019. We used XGBoost due to it producing the highest prediction performance based on our previous work and also because it generates an explainable model[148]. All Albertans aged 18 and over were included in this study. Because the CPSA only has access to certain drug history data (i.e., medication dispensations) and not comorbidity, diagnoses nor lab data, we could not exclude any patients based on comorbidities.

Data Sources

Although a wide range of administrative databases are available in Alberta, few are readily available to professional regulatory agencies like the CPSA, a common issue in many jurisdictions. Indeed, the CPSA currently only has access to prescription drug records in their role as administrator of TPP Alberta. Thus, we limited the datasets used to train the ML model to those accessible by the CPSA. Health regulators in other jurisdictions may have access to additional data to train ML models and create their own specific ML classifier.

To train the ML model, we used data from the Pharmaceutical Information Network (PIN) which has comprehensive information on dispensing records from community pharmacies in Alberta irrespective of coverage status and age[109]. This PIN data was further filtered using Anatomical Therapeutic Chemical classification (ATC) codes to include only those prescription records for which the CPSA has access to[141]. The CPSA receives records of opioid, benzodiazepine and antibiotic dispensations in daily updates from Alberta Health.

To label each instance of an opioid dispensation with an outcome, we linked the prescription data with *1) Population and Vital Statistics Data (VS, Alberta Services)*: sex, age, date of birth, immigration and emigration data within the province, death date and underlying cause of death according to the World Health Organization algorithm using ICD codes (International Statistical Classification of Diseases and Related Health Problems[77]), and *2) Hospitalizations and Emergency Department Visits (NACRS [National Ambulatory Care Reporting System], DAD [Discharge Abstract Database]):* all services, length of stay, diagnoses (up to 25 ICD-10 based diagnoses). Data and coding accuracy are routinely validated both provincially and centrally via the Canadian Institute for Health Information[110]

These linked databases represent a labeled data set used to develop the XGBoost model. Currently, the CPSA does not have access to outcomes data (e.g., DAD, NACRS and VS) nor Physician Claims data. We labelled data equivalent to the PIN data CPSA would receive with DAD, NACRS and VS data to assess a proof-of-concept ML model.

Measures and Outcomes

In this study, the unit of analysis was at the opioid dispensation level such that each opioid dispensation was treated as independent and served as a potential instance to predict our outcome. We chose this level of analysis to be consistent with others[37], to have more data to train the ML model, and to accurately represent clinical use in the real world in which a health regulator's PDMP may want to assess the risk for each instance rather than a single or random dispensation. We identified opioid dispensations from the PIN file using Anatomical Therapeutic Chemical classification (ATC) codes[80] (Table 6.4).

Our outcome was a composite of a drug related emergency department (ED) visit, hospitalization or death within 30-days of an opioid dispensation based on ICD-10 codes from the linked databases (Table 6.5)[23]. One month risk windows are commonly used by health systems for risk assessments[82]. Follow-up and predictions started after each opioid dispensation.

Rare outcomes are common in clinical prediction model development and we anticipated this study to be no different as our training data would be class imbalanced[83]. To

address this, we used the frequently used class weightage method which does not alter the data distribution and instead, increases the importance of the positive class (instances that led to the outcome)[84].

Predictors and ML Methods

All candidate predictors were obtained from the PIN dataset and were either derived from the literature[24,73] or directly incorporated from the data unmodified (Table 6.4). These features included demographics (age, sex, Forward Sortation Index [FSI] from postal codes[86], income), drug utilization (ATC codes, oral morphine equivalents[88], concurrent use with benzodiazepines, number of dispensations, number of unique molecules) and health care utilization (number of opioid prescribers and pharmacies). We used data 30 days before each opioid dispensation to measure each predictor (Figure 6.6).

We used XGBoost to train our ML model. The opioid dispensations in 2018 were included in the development set while those in 2019 were used for validation. We performed k=5 cross fold validation in the development set to tune hyperparameters. With XGBoost, we tuned for tree height, number of trees and weight scaling to address class imbalance[95,120].

Two validations sets were defined from the 2019 dispensations. The first used all of the 2019 dispensations thus, is not a true external independent set as is usually seen in the ML prediction model research literature. In this validation set, labeled as the "entire validation set", participants in the development set were also in the validation set, as this represents the real-world scenario that health regulators work in where patient instances are repeatedly encountered; indeed, it is ideal to develop a ML prediction model trained on data from the population it will be deployed in. In the second validation set, an external set with out of sample instances was defined in which participants in this validation set were not included in the development set, representing an "external" validation set which was a subset of the entire validation set (Figure 6.1). Most of our analysis was done on the entire validation set because health regulators frequently monitor dispensations with repeated encounters making it impractical to report performance metrics on an external validation set with out of sample instances.

Analysis and Prediction Evaluation

We first described characteristics and outcome event rates in the development vs validation group and between those who experienced the outcome and those who did not using chi-square and t-tests. This descriptive analysis was done at the patient level in which each patient could be represented by multiple instances of opioid dispensations. We included information on age, sex, final sample size and healthcare utilization.

The entire validation set was used to evaluate our ML model's prediction performance using metrics that are commonly applied to clinical prediction models[18-20,38]. As is done in many ML prediction studies[17], we assessed our XGBoost model's discrimination performance by estimating the area under the receiver operating characteristics curve (AUROC). For binary classification studies like ours, AUROC curves correspond to c-statistics which are a measure of model discrimination performance, the extent to which a model predicts a higher probability of an outcome among participants who actually had the outcome compared to those who did not[19]. A precision-recall curve (PRC) was also included[89]. For comparison, we estimated the AUROC for the external validation set representing discrimination performance in new patients.

We also provided a calibration plot[19] for the ML model; calibration is considered an important property of any prediction model and reflects the extent to which predicted values align with observed values and is most often illustrated by a plot of observed vs predicted[18,19]. As well, we added a negative predictive value (NPV) vs. predicted risk plot to highlight the relationship between low predicted risk and true negatives (those who did not experience the outcome).

From here, we reported two methods for health regulators to assess the clinical utility of the ML model. The first involved ranking our predicted risks, as others have done[43,66], by categorizing them into percentiles (e.g., deciles) or keeping them in absolute numbers (e.g., top 10 highest risk dispenses). At each of these category cut-off points, we reported prediction performance metrics. These included positive likelihood ratios (LR+)[20,145], true/false positives, true/false negatives, and positive predictive values (PPV, equivalent to post-test probability). These metrics were also reported on the actual thresholds of predicted risk outputted by the

ML classifier. We carried out this analysis on all the data from the validation set and measured these metrics at the end of 2019.

In the second method, we performed a decision curve analysis[21] in which the net benefit of our ML model is compared against two alternatives, namely, intervening on all opioid dispensations or on none, using the entire range of probability threshold cut-off points. This comparison is done by using predicted probabilities from our ML model and comparing them against a probability threshold to aid a decision.

Thus, if a health regulator such as the CPSA is interested in intervening, for example, on the top 1 percentile of predicted risk or top 10 highest risk predicted opioid dispenses, then method 1 could be considered. Alternatively, if they want to intervene on opioid dispensations above a certain predicted risk threshold, method 2 could be informative. Either way, the amount of workload created by identifying high risk dispenses is an important factor for the health regulator when applying this ML classifier; any interventions aided by ML prediction should only increase workload to a manageable extent.

Because we are using Alberta data, we simulated predictions to view the capabilities of our ML model if deployed into CPSA workflow. These included predictions measured daily and weekly that progressively excluded participants once they were already flagged as high risk. Filtering out previously flagged patients represents a more realistic scenario for any health regulator as it is not practical to repeatedly identify the same high-risk patients. For this simulation, previously flagged participants were excluded for the entire year keeping in mind that a health regulator could exclude patients on a monthly or quarterly basis. For comparison, we reported the results of simulating and stratifying predictions using percentiles by not progressively excluding participants previously flagged as high risk. We also simulated the number of 30-day events per 100 daily dispenses stratified by percentiles of risk. Considering predicted risk thresholds and workload, we will report how many high-risk dispenses the CPSA would have to consider based on threshold cut-off points.

Since ML models do not estimate an interpretable quantity relating predictors to outcomes, it is not appropriate to summarize that relationship with a single parameter.

Instead, the impact of individual predictors can be summarized using "variable importance", which is a rank-ordering of variables that are most important for the ML model's prediction performance[50]; variable importance does not have a causal or statistical meaning. To address interpretability[59] of our ML model, we reported feature importance[50] and feature impact using SHAP (Shapley Additive Explanations) value plots[91,92], which will give the CPSA some insight on how the ML model was influenced in its predictions.

We briefly describe the general format of an electronic dashboard which health regulators could use in their opioid stewardship program.

Cost Analysis

We performed a real-world simulation of cost savings of a ML model assisted PDMP using TPP Alberta as an example and reporting on the second quarter (Q2) of our 2019 data as an illustration. Multiplying the Resource Intensity Weight (RIW)[149,150] associated with each admission in the administrative databases by the Cost of a Standard Hospital Stay (CSHS) metric[151] developed by CIHI[57], we calculated the cost of each hospitalization or ED visit related to our outcome. This is a validated method used by researchers[152,153] and CIHI[57,58] to estimate costs in the health care system.

We estimated a summary cost of all admissions and ED visits related to our outcome for Q2 2019. Study participants were progressively excluded once they experienced our defined outcome for the remainder of the quarter. To calculate costs associated with our predictions, we added up the costs of all the drug related hospitalizations and ED visits in the quarter starting at the first positive instance of an opioid dispensation. As done in other studies[43], we used the entire range of percentiles of predicted risk to graphically illustrate health care costs associated with our true positive predictions. For each of the top 1, 5, and 10 percentile categories of predicted risk, we described potential savings to the health system by assessing cost reductions stratified by a range of intervention success rates and costs as others have done[125]. For this study, we only considered a general point of care intervention such as a physician follow-up because we could reasonably estimate a range of costs. All dollar figures are in Canadian currency.

We did not anticipate any missing data in our study because TPP Alberta PIN data involves full capture of all information.

This study followed the TRIPOD[68] and other guidelines[11,17] specific to ML projects. All analyses were done using Python (version 3.6.8, Python Software Foundation), SciKit Learn[94] (version 0.23.2), SHAP[92] (version 0.35), XGBoost[95] (version 0.90), Pandas[96] (version 1.0.5) and STATA/MP V.15.1 (StataCorp). This study received ethics approval from the University of Alberta ethics board (Pro00083807).

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.

Data availability

The data that support the findings of this study are available from Alberta Health but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. However, administrative health data can be accessed from Alberta Health by following defined research protocols and confidentiality agreements.

Results

A total of 853,324 participants were included in this study representing 6,181,025 opioid dispenses during 2018-2019. During the same time period, 145,016 outcome events (2.3%) occurred. Dispenses in 2018 and 2019 comprised the development and validation sets, respectively (Figure 6.1). The validation set had 77,326 opioid dispenses with positive instances (2.6%) representing the pre-test probability of the outcome and averaged around 8,241 (SD=2,423) opioid dispenses per day (Figure 6.7). Characteristics were comparable between those in the development and validation sets (Table 6.6) while differences were noted between those who experienced the outcome and those who did not (Table 6.7), as is expected.

Using the entire validation set, we estimated the AUROC for the XGBoost classifier to be 0.82 and an area under the PRC of 0.13 (Figure 6.8). The corresponding values for the external validation set were 0.89 and 0.21, respectively. As for the calibration plot, the observed and predicted risks were not aligned and showed a consistent overestimation of risk with a substantial fraction of instances predicted as low risk (Figure 6.9). Low predicted risks were accompanied by fewer actual outcomes highlighting higher NPVs at lower predicted risks (Figure 6.10).

After we ranked and grouped our predicted risks at the end of 2019, the categories with the highest risk predictions had the highest PPVs (and post-test probabilities) and LR+'s. The top 0.1 percentile of predicted risk had a LR+ of 28.7 which translated to a post-test probability of 43.1% compared to the pre-test probability of 2.6% (Table 6.1). Similar results were observed with the top 10 high risk dispenses reporting a PPV of 0.9 and LR+ of 341 (Table 6.1). There was also an increase in LR+ and PPV as the threshold of predicted risk increased (Table 6.8).

When we performed the decision curve analysis across the entire range of threshold probabilities, the XGBoost classifier provided no more value than if all or none of the participants were considered high risk (Figure 6.2).

In our simulations, the predictions classified weekly had higher LR+ in both the highest risk dispenses and percentiles of predicted risk as compared to predictions classified daily. Measured weekly, we reported a LR+ of 20.55 in the top 20 dispenses corresponding to a PPV (post-test probability) of 0.20. This is in comparison to daily measured predictions which had LR+ and PPV of 6.38 and .12, respectively (Table 6.2). The same trend occurred when we assessed weekly vs daily predictions using top percentiles of predicted risk (Table 6.9). By progressively excluding previously flagged participants and measuring the top 20 riskiest dispenses, the total number of 30-day events decreased as the year progressed (Figure 6.3A). When we assessed performance in the top percentiles of predicted risk and not excluding previously flagged participants, LR+'s were higher (Table 6.10) as compared to the results of excluding previously flagged participants. Considering actual probability thresholds and

workload, as the predicted probability threshold cut-off point increased, the number of flagged dispenses decreased for predictions measured daily and weekly (Table 6.3). Also, higher thresholds were connected to higher LR+, with weekly measurements being more informative than daily ones (Table 6.3).

When we simulated based on top percentiles of predicted risk and events per 100 daily dispenses, the highest 1 percentile of predicted risk had higher event rates than lower predicted risk percentiles, including the baseline risk (pre-test probability of around 2.6%; Figure 6.3B).

With respect to ML interpretability, previous opioid dispensations and age were ranked highest in variable importance (Figure 6.4). Higher number of previous opioid dispenses and younger age were suggestive of higher risk of a 30-day event (Figure 6.11).

We included the outputs from our ML model within an electronic dashboard which allows the CPSA to identify patients at high risk for a 30-day event and decide what, if any, interventions to initiate. This patient centric view of the data allows the CPSA to filter an aggregated patient list based on ML risk prediction scores, age, and drug utilization characteristics (e.g., oral morphine equivalents, benzodiazepine dispenses). By selecting a patient from the resulting list, users can reveal that patient's individual SHAP values related to their risk prediction score for the associated opioid dispense. Full dispensation profiles are also available. The relevant SHAP values displayed are sorted from highest risk factor to highest protective factor in a split bar chart, which makes it very easy for the CPSA user to see what factors contribute the most to patients' overall risk score. Hovering over one of the bars gives further details on the underlying feature values that contributed to the risk score to aid ML prediction interpretation. To help with use of the dashboards, we also included a page with general use instructions, and the list of all features along with a brief description for each. It is also possible for users to search for a particular patient even if they are not among the highest of predicted risk.

Cost analysis

In 2019 Q2, we estimated around $57.4 million was spent on opioid related admissions signifying a substantial space for cost savings. Around $12 million (21%) represents admission costs associated with our true positive predictions (Figure 6.5A). Cost savings diminished as the cost of intervention increased across all intervention success rates. In the top 1 percentile of predicted risks, intervention success rates above 50% were associated with cost savings across the range of intervention costs, while those below 50% added costs to the health system at higher intervention costs (Figure 6.5B). In the top 5 and 10 percentiles of predicted risk, the same trend was seen except that all intervention success rates added costs to the system at higher intervention costs (Figure 12).

Discussion

In this study, we created an XGBoost ML classifier for PDMPs to assess risk from prescribed opioids using the CPSA as an example of a health regulator. We presented 2 analytic options for health regulators to implement ML decision support into an opioid stewardship workflow namely, acting on the highest ranked predictions or on probability threshold cut-offs. Discrimination, calibration and net benefit analysis are important aspects in determining the clinical utility of a prediction model[21]. Although our model had strong discrimination performance, it did not calibrate well. Our net benefit analysis suggests that using probability thresholds may not lead to an informative decision aid. However, there may be some value in ranking predictions as some LR+'s led to conclusive changes from pre to post test probabilities[20] with weekly predictions being more informative than daily ones. Our findings also point to merit in acting on ranked predictions to reduce costs to the health system depending on accepted intervention success rates and costs. Miscalibration and uninformative net benefit analysis could partly be explained by the limited data available to the CPSA. From previous work, it is well known that having more data could better leverage ML prediction capabilities[16,34] thus making the case to increase data access and permissions for the CPSA and regulators in general, especially if intervening on probability thresholds is important.

Using this option of ranking predictions, health regulators could implement ML prediction as a decision aid to potentially intervene on high-risk opioid dispenses and reduce

health care costs. The electronic dashboard we described can be adjusted according to acceptable workloads by varying the number of identified high risk opioid dispenses and the variable importance SHAP values can provide some understanding of why the ML classifier flagged these high-risk instances. Regulators could also progressively exclude previously flagged patients yearly, as we did, quarterly or other based on workload capacity.

To date, we are unaware of other Canadian jurisdictions which have implemented or studied ML prediction to aid PDMPs in reducing risk from opioids and to reduce health costs. Although trained on the limited data available to the CPSA, the discrimination performance of the ML model in this study was comparable to results from our previous work and that of others[37]. Furthermore, the prediction performance using ranked predictions was comparable to that in our previous work, in which we had access to more types of training data than the CPSA.

Limitations in our study are mainly due to data issues. Our training dataset does not account for the risk associated with non-prescription opioid use. The CPSA's limited data access did not allow for any exclusion criteria in our study population thus, we were not able to exclude any participants based on comorbidity or other (e.g., cancer, palliative care) histories. This issue could be mitigated by allowing the CPSA increased access to other administrative datasets storing social factors and comorbidity data for training ML models. Nevertheless, we demonstrated informative predictions in the higher risk dispenses and even with our limited data, we were still able to demonstrate cost savings using our ML predictions. Our ML classifier was trained only on Alberta data, which may not be generalizable to other jurisdictions. However, the ML process makes it easy to train models using population specific data.

This study considered a particular regulator's perspective with respect to clinical utility, workload and ML interpretability when predicting opioid related outcomes. This approach can be applied to other health regulators with similar opioid stewardship mandates. Training on data representing specific populations and regular retraining to improve prediction performance over time are among the benefits of the ML process. Our XGBoost classifier gives health regulators options for interventions based on workload capacity; any potential

interventions on opioid dispenses can be identified at the patient level in which opioid use could be further assessed. Improved data access could improve prediction performance, leading to more effective opioid stewardship and further reductions in health care costs.

Table 6.1. XGBoost prediction metrics measured at the end of 2019.

		Threshold	TP	FP	FN	TN	PPV*	Se	Sp	LR+
Top dispenses	**10**	0.982	9	1	77,317	2,931,070	0.900	0.0001	1.0000	341.1
	50	0.978	46	4	77,280	2,931,067	0.920	0.0006	1.0000	435.9
	100	0.976	79	21	77,247	2,931,050	0.790	0.0010	1.0000	142.6
	500	0.965	250	250	77,076	2,930,821	0.500	0.0032	0.9999	37.9
	1000	0.960	498	502	76,828	2,930,569	0.500	0.0064	0.9998	37.6
	5000	0.937	2,014	2,986	75,312	2,928,082	0.400	0.0260	0.9990	25.5
	10000	0.920	3,462	6,538	73,864	2,924,533	0.350	0.0448	0.9978	20.1
	50000	0.856	10,779	39,221	66,547	2,891,850	0.216	0.1394	0.9866	10.4
	100000	0.812	17,238	82,762	60,088	2,848,309	0.172	0.2229	0.9718	7.9
Top percentile of predicted risk	**0.01**	0.970	183	118	77,143	2,930,953	0.608	0.0024	1.0000	55.5
	0.1	0.946	1,295	1,713	76,031	2,929,358	0.431	0.0167	0.9994	28.7
	1	0.882	7,559	22,525	69,767	2,908,546	0.251	0.0978	0.9923	12.7
	5	0.780	22,175	128,245	55,151	2,802,826	0.147	0.2868	0.9562	6.6
	10	0.711	32,182	268,658	45,144	2,662,413	0.107	0.4162	0.9083	4.5
	25	0.569	51,526	700,573	25,800	2,230,498	0.069	0.6663	0.7610	2.8
	50	0.352	69,659	1,434,539	7,667	1,496,532	0.046	0.9008	0.5106	1.8
	75	0.149	75,719	2,180,579	1,607	750,492	0.034	0.9792	0.2560	1.3
	90	0.072	76,968	2,630,589	358	300,482	0.028	0.9954	0.1025	1.1

TP: true positives; FP: false positives; FN: false negatives; TN: true negatives; Se: sensitivity; Sp: specificity; LR+: positive likelihood ratio; PPV: positive predictive value (post-test probability)

*Compared with pre-test probability of 2.6% based on prevalence.

Table 6.2. Prediction metrics simulated using daily and weekly measurements stratified by top dispenses using 2019 data. Participants were progressively excluded for 1 year if previously flagged as high risk.

	Top dispenses	Threshold	TP	FN	FP	TN	PPV*	NPV	Se	Sp	LR+
Measured daily	10	0.83	1	127	10	7506	0.12	0.98	0.01	1.00	8.00
	20	0.78	2	105	19	7108	0.09	0.99	0.02	1.00	6.38
	50	0.67	3	65	49	6167	0.06	0.99	0.04	0.99	5.63
	100	0.54	4	41	99	5187	0.04	0.99	0.09	0.98	4.74
	200	0.40	5	24	203	4204	0.02	0.99	0.17	0.95	3.76
	500	0.22	6	9	627	2517	0.01	1.00	0.41	0.80	2.07
	1000	0.09	7	3	1185	929	0.01	1.00	0.69	0.44	1.24
Measured weekly**	10	0.92	3	510	8	39838	0.24	0.99	0.01	1.00	24.66
	20	0.90	4	488	17	39701	0.20	0.99	0.01	1.00	20.55
	50	0.86	8	444	43	39285	0.15	0.99	0.02	1.00	15.25
	100	0.81	11	395	90	38627	0.11	0.99	0.03	1.00	11.64
	200	0.74	16	333	186	37341	0.08	0.99	0.04	1.00	9.06
	500	0.61	25	218	496	33860	0.05	0.99	0.10	0.99	7.20
	1000	0.46	31	145	1018	29377	0.03	1.00	0.18	0.97	5.24

Note: Based on average daily and weekly values to prevent daily and weekly fluctuations of dispenses.

Threshold: predicted probability threshold; TP: true positives; FP: false positives; FN: false negatives; TN: true negatives; Se: sensitivity; Sp: specificity; LR+: positive likelihood ratio; PPV: positive predictive value (post-test probability)

*Compared with pre-test probability of 2.6% based on prevalence.

**The highest predicted probability of the week was used.

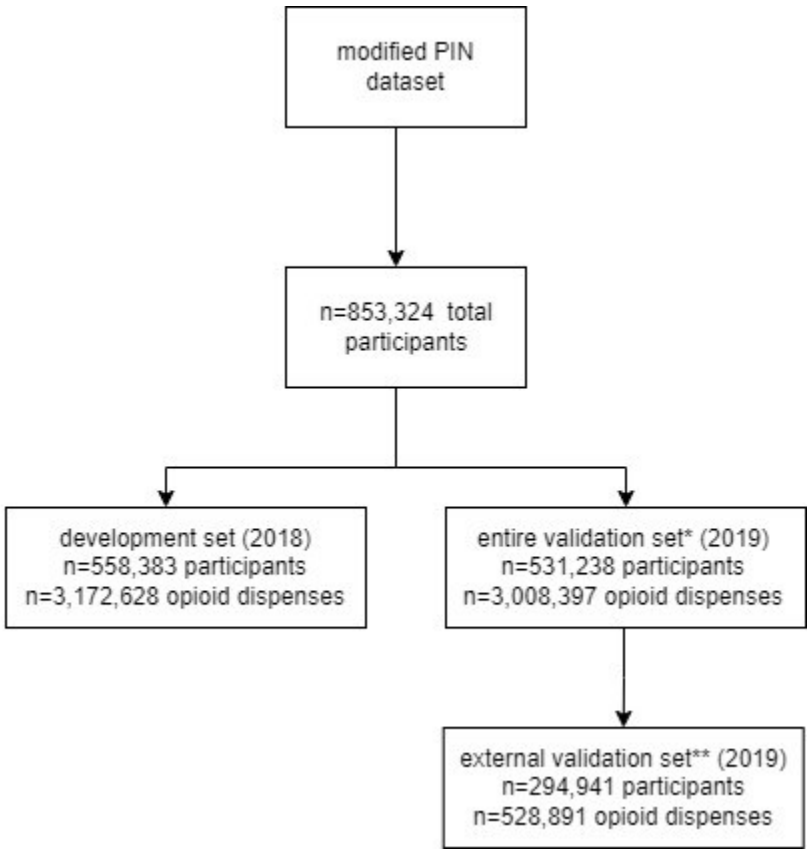
Table 6.3. Prediction metrics simulated using daily and weekly measurements stratified by absolute probability thresholds using 2019 data. Participants were progressively excluded for 1 year if previously flagged as high risk.

	Threshold	No. of flagged dispenses	TP	FN	FP	TN	PPV*	NPV	Se	Sp	LR+
Measured daily	0	1457	7	0	1450	0	0.00	(--)	1.00	0.00	1.00
	0.1	1009	7	1	1002	811	0.01	1.00	0.90	0.45	1.63
	0.2	569	6	3	563	1825	0.01	1.00	0.69	0.76	2.93
	0.3	316	5	6	311	2578	0.02	1.00	0.49	0.89	4.52
	0.4	217	5	9	212	3126	0.02	1.00	0.35	0.94	5.54
	0.5	150	4	13	146	3692	0.03	1.00	0.25	0.96	6.51
	0.6	98	4	23	94	4479	0.04	0.99	0.14	0.98	6.76
	0.7	53	3	52	51	5676	0.05	0.99	0.05	0.99	5.82
	0.8	22	2	91	21	6825	0.08	0.99	0.02	1.00	6.47
	0.9	4	1	154	3	7757	0.15	0.98	0.00	1.00	8.96
	1	0	0	212	0	8029	(--)	0.97	0.00	1.00	(--)
Measured weekly**	0	10198	50	0	10148	0	0.00	(--)	1.00	0.00	1.00
	0.1	7060	48	5	7013	5529	0.01	1.00	0.90	0.44	1.62
	0.2	3984	42	18	3941	12411	0.01	1.00	0.70	0.76	2.91
	0.3	2212	37	36	2175	17483	0.02	1.00	0.51	0.89	4.58
	0.4	1516	33	54	1483	21083	0.02	1.00	0.38	0.93	5.79
	0.5	1051	31	79	1020	24570	0.03	1.00	0.28	0.96	6.97
	0.6	683	26	123	656	28539	0.04	1.00	0.18	0.98	7.82
	0.7	374	20	209	355	33097	0.05	0.99	0.09	0.99	8.10
	0.8	157	13	319	144	36904	0.08	0.99	0.04	1.00	9.77
	0.9	29	4	465	24	39418	0.15	0.99	0.01	1.00	15.18
	1	0	0	558	0	39988	(--)	0.99	0.00	1.00	(--)

Note: Based on average daily and weekly values to prevent daily and weekly fluctuations of dispenses; number of flagged dispenses represents a potential CPSA workload.

Threshold: predicted probability threshold; TP: true positives; FP: false positives; FN: false negatives; TN: true negatives; Se: sensitivity; Sp: specificity; LR+: positive likelihood ratio; PPV: positive predictive value (post-test probability)
*Compared with pre-test probability of 2.6% based on prevalence.
**The highest predicted probability of the week was used.

Figure 6.1. Study participant flow diagram. Note: PIN=Pharmaceutical Information Network



*Participants could be in both the development and validation set; the analysis was conducted on this validation set

**Participants in this validation set are not included in the development set

Figure 6.2. Decision curve analysis. Across most of the range of threshold probabilities, the XGBoost classifier had a lower net benefit than if none of the opioid dispenses were intervened on. Thus, acting on predicted probability thresholds for interventions may not be informative nor appropriate.

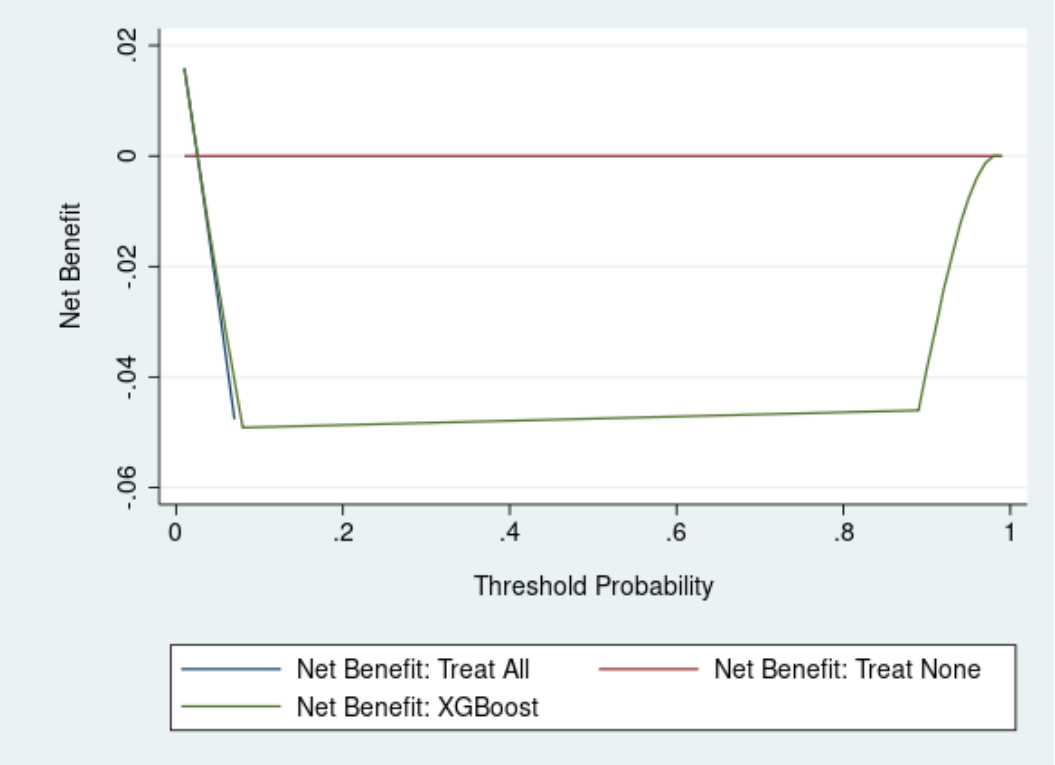
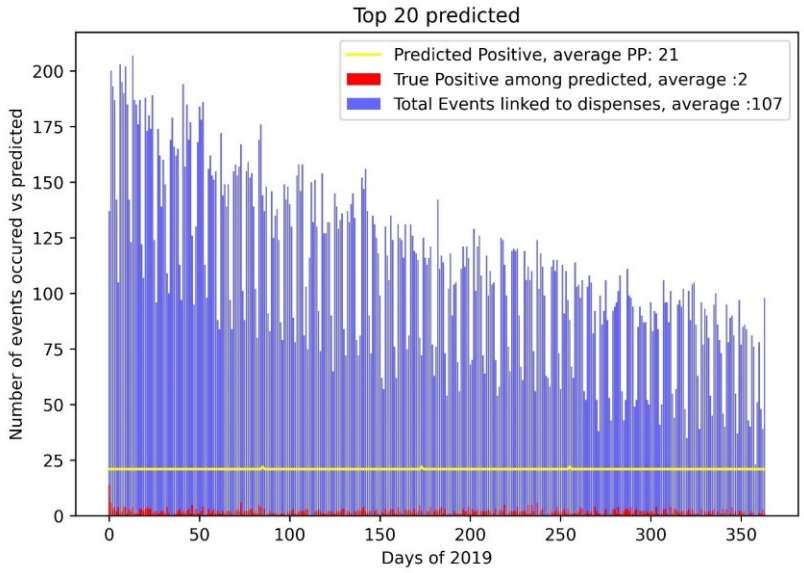


Figure 6.3. Simulation of predicting the top 20 riskiest opioid dispenses measured daily by progressively excluding participants previously flagged as high risk. The yellow line, what we predicted, represents a workload that the CPSA would have to consider (A); XGBoost classifier predicting daily risks in a simulation for the College of Physicians and Surgeons of Alberta stratified by top percentile categories of risk. Base risk is around 2.6% and represents the pre-test probability (B).

A)



B)

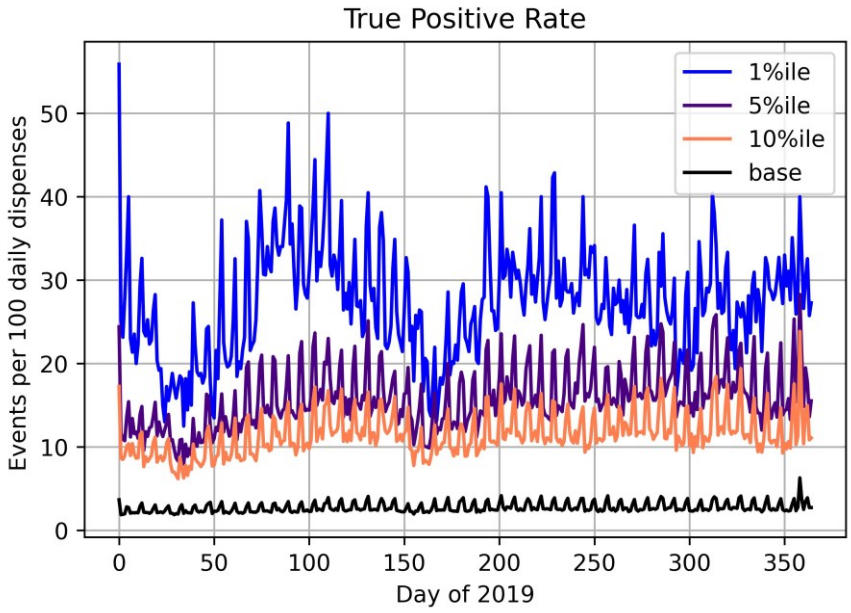


Figure 6.4. Variable importance for the XGBoost classifier. Variable importance bears no statistical meaning in terms of association.

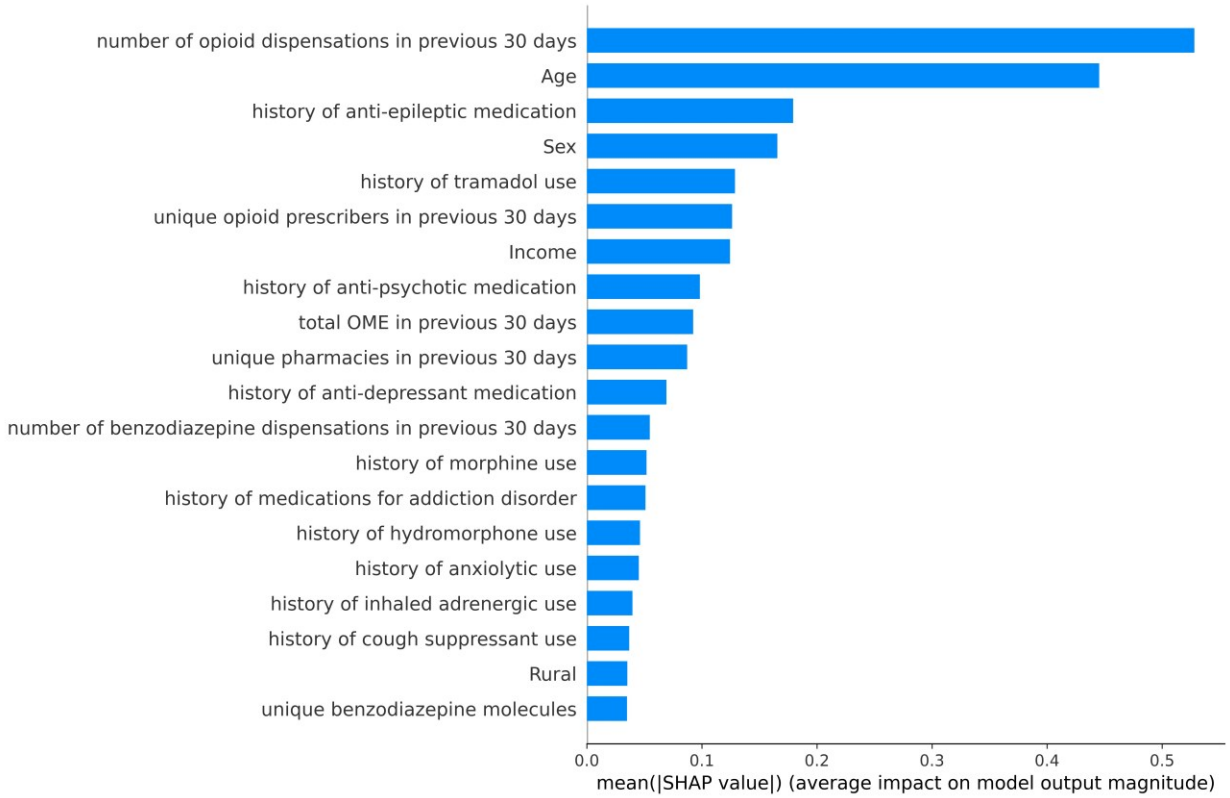
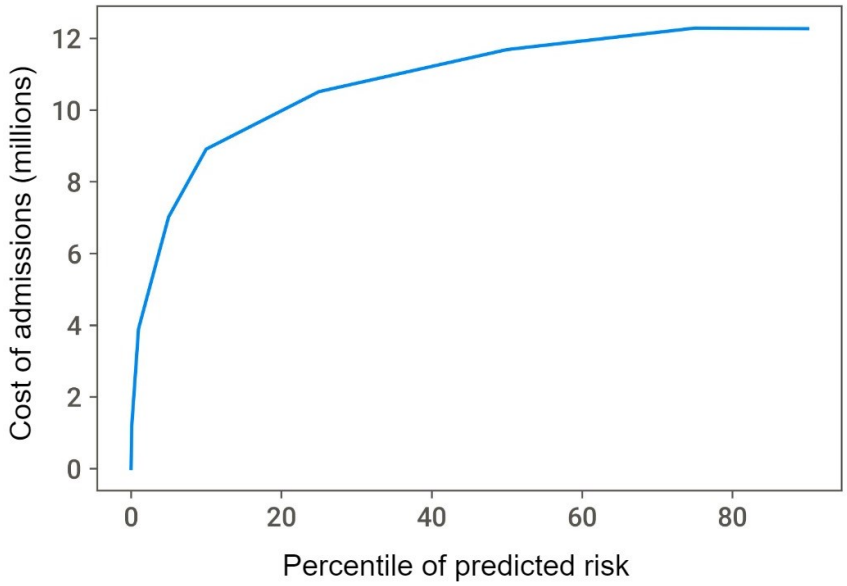
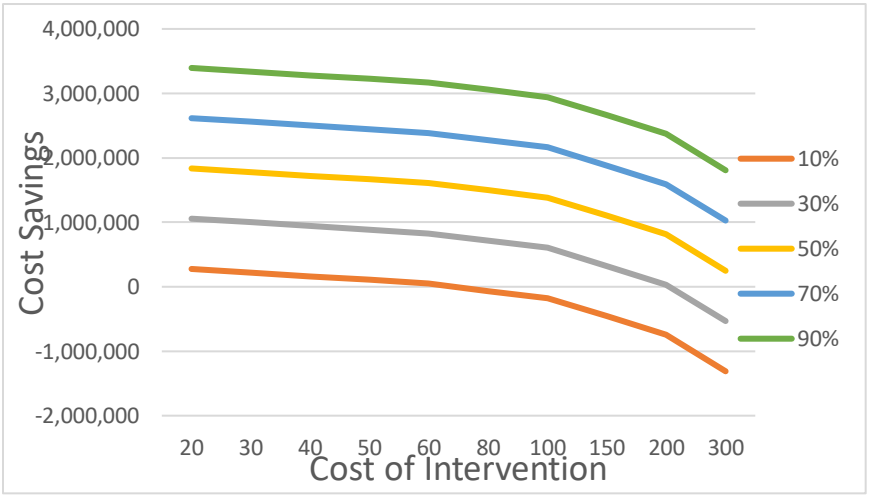


Figure 6.5. Cost of admissions (hospitalizations and emergency department visits) using predictions ranked by percentiles (percentile categories are the cut-off points) for data in 2019 Quarter 2. Costs are associated with only true positive predictions and represent the maximum possible savings the machine learning classifier will realize at the given percentile threshold of prediction based on daily classifications by a health regulator (A); Cost savings and cost of interventions stratified by intervention success rates for predictions ranked in the top 1 percentile for 2019 Quarter 2 (B). All dollar amounts are in Canadian currency.

(A)



(B)



Appendix to Chapter 6

Table 6.4. Anatomical Therapeutic Chemical classification of opioid molecules used for this study and candidate predictors used to develop the XGBoost classifier.

Category (data source)	Description
ATC codes used to identify opioids from modified PIN data	N01AH01, N01AH03, N01AH06, N07BC01, N07BC02, N07BC51, R05DA03, R05DA04, R05DA09, R05DA20, N02A
Opioid molecules used in this study (modified PIN)	alfentanil, butorphanol, codeine, diamorphine, fentanyl, hydrocodone, hydromorphone, meperidine, morphine, oxycodone, oxymorphone, pentazocine, sufentanil, tapentadol, tramadol
Demographic information (modified PIN)	age, sex, postal codes, mean income
Drug utilization history (modified PIN)	drug dispenses in past 30 days using 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
Healthcare utilization (modified PIN)	Number of opioid prescribers and pharmacies

Note: ATC- Anatomical Therapeutic Chemical classification (https://www.whocc.no/atc_ddd_index); modified PIN- Pharmaceutical Information Network data modified to align with data access granted to the College of Physicians and Surgeons of Alberta.

Table 6.5. Diagnostic codes used to identify the defined study outcome from emergency department 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

Table 6.6. Characteristics of study participants (n=853,324) in the development and validation sets.

Characteristic	Number in development set *	Percent	Number in validation set*	Percent
Number of participants	558,383	100.0	531,238	100.0
Age:				
Mean (SD)	50.7 (17.6)	(--)	51.0 (17.7)	(--)
18-45	226,648	40.6	213,456	40.2
45-65	212,760	38.1	200,816	37.8
>65	118,975	21.3	116,966	22.00
Male	265,954	54.0	243,639	45.9
Female	301,429	46.0	287,599	51.1
Rural	88,783	15.9	83,713	15.8
Urban	469,600	84.1	447,525	84.2
Income:				
<75k	83,010	14.9	77,901	14.7
75k-100k	286,509	51.3	270,397	50.9
100k-171k	175,837	31.5	171,038	32.2
>171k	13,027	2.3	11,902	2.2
Number of unique physicians visited in the 30-days prior to opioid dispensation:				
1 to 3	532,546	95.4	503,745	94.8
>3	25,837	4.6	27,493	5.2

*Unless otherwise indicated

Note: p<0.001 for all comparisons due to large sample sizes except rural/urban (p=0.04); participants can be in either or both categories.

Table 6.7. Characteristics of study participants (n=853,324) according to outcome status.

Characteristic	Number with event*	Percent	Number without event*	Percent
Number of participants	14,916	100.0	852,065	100.0
Age:				
Mean (SD)	45.8 (15.2)	(--)	49.8 (17.8)	(--)
18-45	7,689	51.5	367,777	43.2
45-65	5,677	38.1	310,897	36.5
>65	1,550	10.4	173,391	20.3
Male	8,191	54.9	394,935	46.3
Female	6,725	45.1	457,130	53.4
Rural	3,536	23.7	128,621	15.1
Urban	11,380	76.3	723,444	84.9
Income:				
<75k	4,104	27.5	120,805	14.2
75k-100k	8,069	54.1	431,444	50.6
100k-171k	2,492	16.7	279,699	32.8
>171k	251	1.7	20,117	2.4
Number of unique physicians visited in the 30-days prior to opioid dispensation:				
1 to 3	10,231	68.6	823,167	96.6
>3	4,685	31.4	28,898	3.4

*Unless otherwise indicated.

Note: $p<0.001$ for all comparisons between event and no event; participants can be in either or both categories because some instances lead to an event while others do not within the same participant.

Table 6.8. XGBoost prediction performance metrics based on threshold of predicted risk measured at the end of 2019.

Threshold of predicted risk	TP	FN	FP	TN	PPV*	Se	Sp	LR+
0	77,326	0	2,931,071	0	0.026	1.000	0.000	1.00
0.1	76,623	703	2,440,890	490,181	0.030	0.991	0.167	1.19
0.2	74,568	2,758	1,927,842	1,003,229	0.037	0.964	0.342	1.47
0.3	71,536	5,790	1,573,273	1,357,798	0.043	0.925	0.463	1.72
0.4	66,987	10,339	1,288,735	1,642,336	0.049	0.866	0.560	1.97
0.5	58,785	18,541	939,305	1,991,766	0.059	0.760	0.680	2.37
0.6	47,742	29,584	598,938	2,332,133	0.074	0.617	0.796	3.02
0.7	33,725	43,601	293,576	2,637,495	0.103	0.436	0.900	4.35
0.8	19,145	58,181	98,486	2,832,585	0.163	0.248	0.966	7.37
0.9	5,385	71,941	13,419	2,917,652	0.286	0.070	0.995	15.21
0.99	0	77,326	0	2,931,071	(--)	0.000	1.000	(--)

TP: true positives; FP: false positives; FN: false negatives; TN: true negatives; Se: sensitivity; Sp: specificity; LR+: positive likelihood ratio; PPV: positive predictive value, post-test probability

*Compared with pre-test probability of 2.6% based on prevalence.

Table 6.9. Prediction metrics simulated using daily and weekly measurements stratified by percentiles of predicted risk using 2019 data. Participants were progressively excluded for 1 year if previously flagged as high risk.

	Top Percentile of predicted risk	Threshold	TP	FN	FP	TN	PPV*	NPV	Se	Sp	LR+
Measured daily	**0.1**	0.86	1	141	6	7664	0.14	0.98	0.01	1.00	9.14
	1	0.66	3	63	49	6153	0.06	0.99	0.04	0.99	5.68
	5	0.41	5	21	173	4181	0.03	0.99	0.18	0.96	4.66
	10	0.31	5	11	296	3349	0.02	1.00	0.32	0.92	3.95
	25	0.20	6	4	643	2058	0.01	1.00	0.60	0.76	2.52
	50	0.11	7	1	982	1055	0.01	1.00	0.84	0.52	1.73
	75	0.07	7	0	1243	440	0.01	1.00	0.96	0.26	1.30
	90	0.04	7	0	1370	157	0.01	1.00	0.99	0.10	1.10
Measured weekly**	**0.1**	0.89	5	474	24	39576	0.18	0.99	0.01	1.00	18.02
	1	0.71	18	299	240	36518	0.07	0.99	0.06	0.99	8.60
	5	0.46	31	128	992	28623	0.03	1.00	0.20	0.97	5.89
	10	0.34	36	77	1726	23802	0.02	1.00	0.32	0.93	4.74
	25	0.21	44	29	4265	14931	0.01	1.00	0.60	0.78	2.70
	50	0.12	48	9	6657	7815	0.01	1.00	0.84	0.54	1.82
	75	0.07	49	2	8636	3233	0.01	1.00	0.96	0.27	1.32
	90	0.04	49	1	9567	1155	0.01	1.00	0.99	0.11	1.11

Note: Based on average daily and weekly values to prevent daily and weekly fluctuations of dispenses.

Threshold: predicted probability threshold; TP: true positives; FP: false positives; FN: false negatives; TN: true negatives; Se: sensitivity; Sp: specificity; LR+: positive likelihood ratio; PPV: positive predictive value (post-test probability)

*Compared with pre-test probability of 2.6% based on prevalence.

**The highest predicted probability of the week was used.

Table 6.10. Prediction metrics simulated using daily and weekly measurements stratified by top percentiles of predicted risk using 2019 data. Participants were NOT progressively excluded if previously flagged as high risk.

	Top Percentile of predicted risk	Threshold	TP	FN	FP	TN	PPV*	NPV	Se	Sp	LR+
Measured daily	0.1	0.95	4	208	4	8025	0.47	0.97	0.02	1.00	33.78
	1	0.89	19	193	50	7979	0.28	0.98	0.09	0.99	14.38
	5	0.79	56	156	286	7743	0.16	0.98	0.26	0.96	7.44
	10	0.73	82	130	603	7426	0.12	0.98	0.39	0.92	5.13
	25	0.6	133	79	1633	6396	0.08	0.99	0.63	0.80	3.09
	50	0.39	187	25	3724	4305	0.05	0.99	0.88	0.54	1.90
	75	0.18	206	5	5748	2281	0.03	1.00	0.97	0.28	1.36
	90	0.09	210	1	7053	976	0.03	1.00	0.99	0.12	1.13
Measured weekly**	0.1	0.93	11	547	19	39969	0.36	0.99	0.02	1.00	41.03
	1	0.85	56	502	232	39756	0.19	0.99	0.10	0.99	17.23
	5	0.74	149	409	1294	38693	0.10	0.99	0.27	0.97	8.24
	10	0.66	223	335	2704	37283	0.08	0.99	0.40	0.93	5.90
	25	0.5	370	188	7579	32409	0.05	0.99	0.66	0.81	3.50
	50	0.29	483	75	16136	23851	0.03	1.00	0.87	0.60	2.15
	75	0.14	537	21	26759	13229	0.02	1.00	0.96	0.33	1.44
	90	0.08	553	5	34261	5727	0.02	1.00	0.99	0.14	1.16

Note: Based on average daily and weekly values to prevent daily and weekly fluctuations of dispenses.

Threshold: predicted probability threshold; TP: true positives; FP: false positives; FN: false negatives; TN: true negatives; Se: sensitivity; Sp: specificity; LR+: positive likelihood ratio; PPV: positive predictive value (post-test probability)

*Compared with pre-test probability of 2.6% based on prevalence.

**The highest predicted probability of the week was used.

Figure 6.6. Schematic of study design and feature generation.

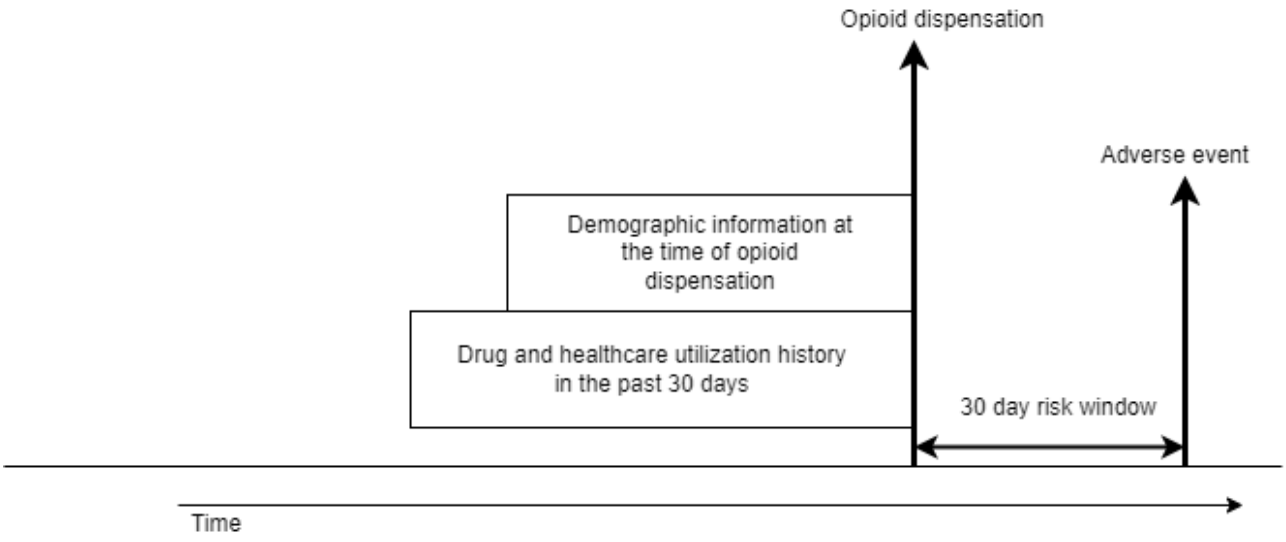


Figure 6.7. The number of opioid dispenses per day during 2019 (average of 8241 dispenses per day) and the number of dispenses which resulted in our defined outcome (average of 212 or around 2.6% of dispenses which led to an event).

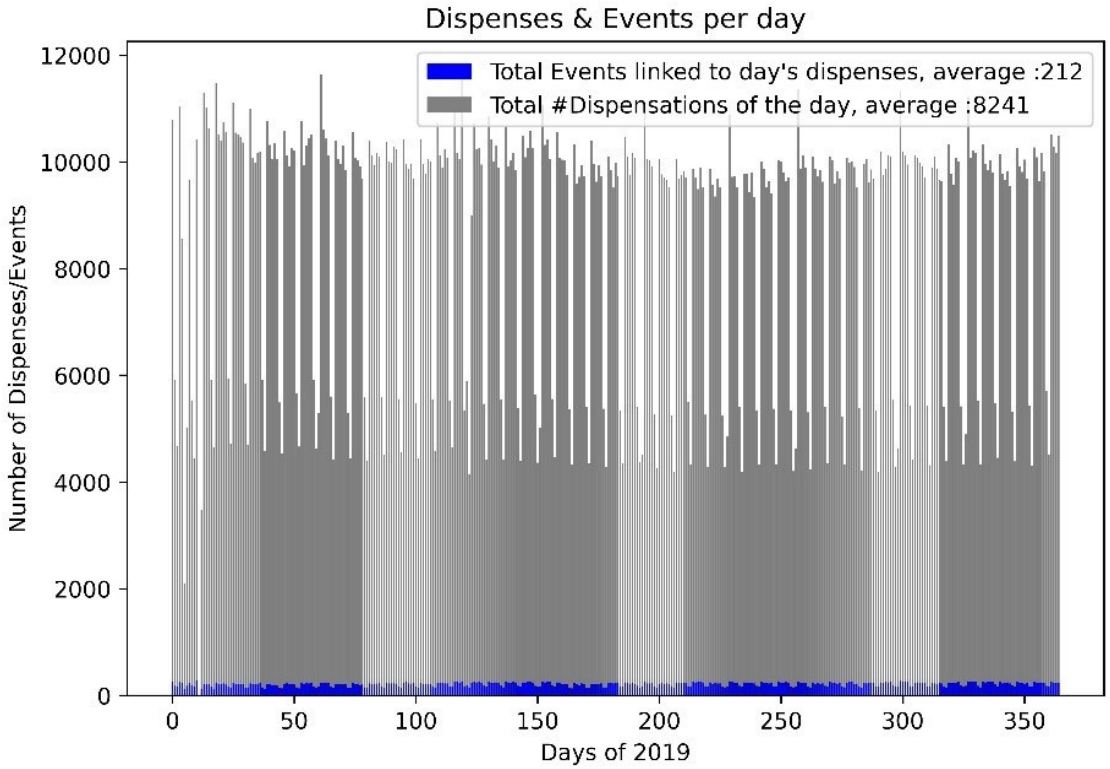
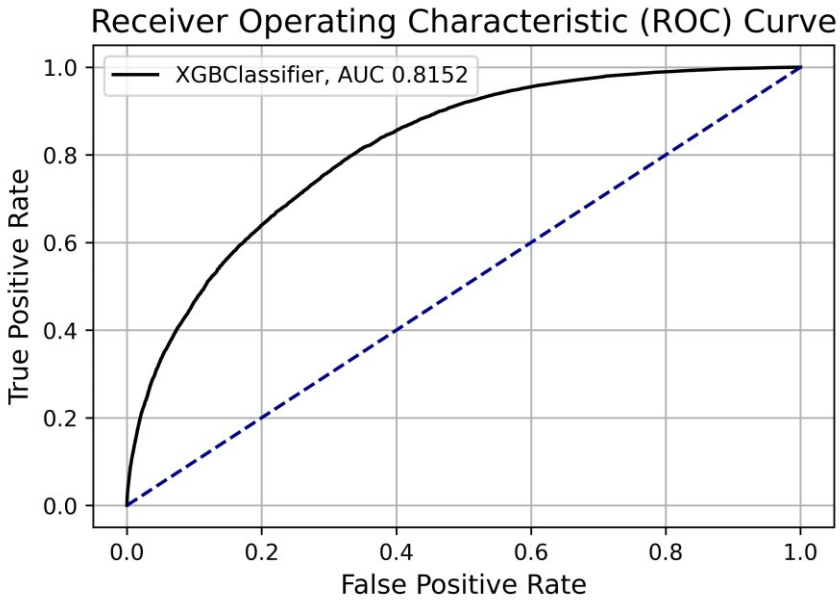


Figure 6.8. Discrimination performance (A) and precision-recall curve(B) of our XGBoost classifier using the entire 2019 validation data.

(A)



(B)

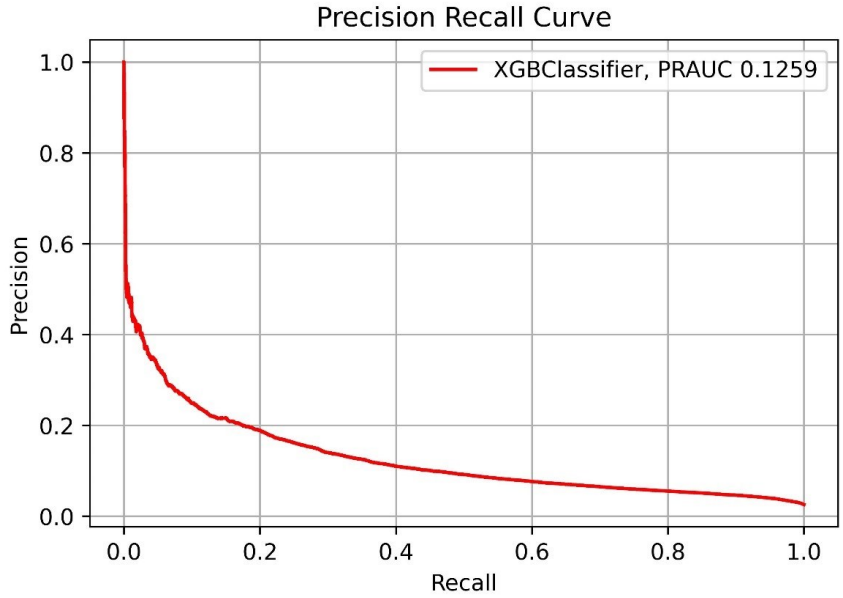


Figure 6.9. Calibration plot of the XGBoost classifier using 2019 data.

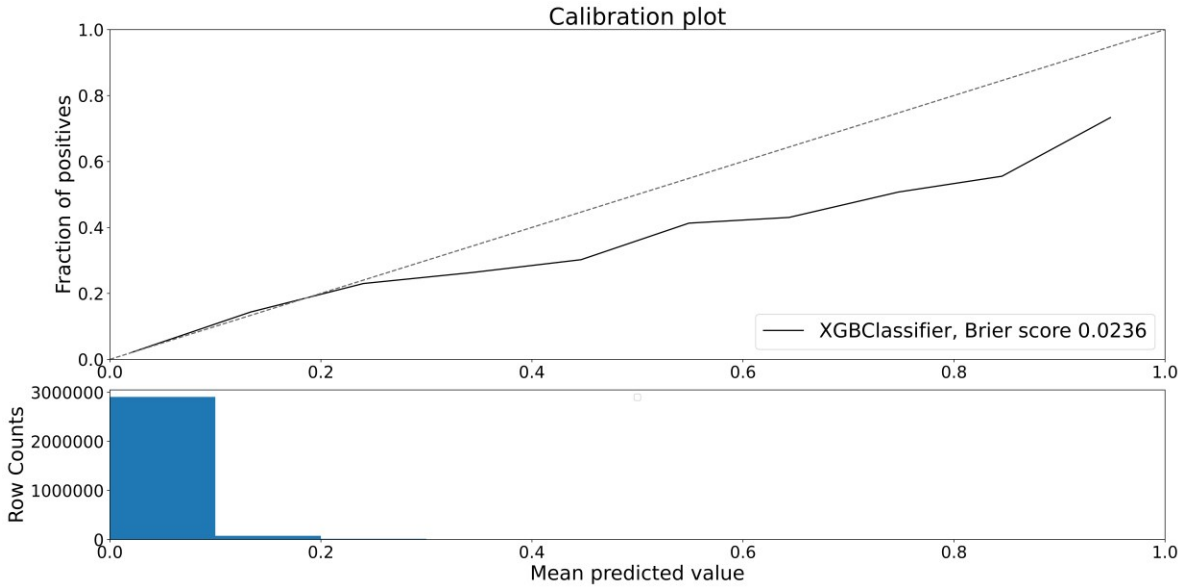
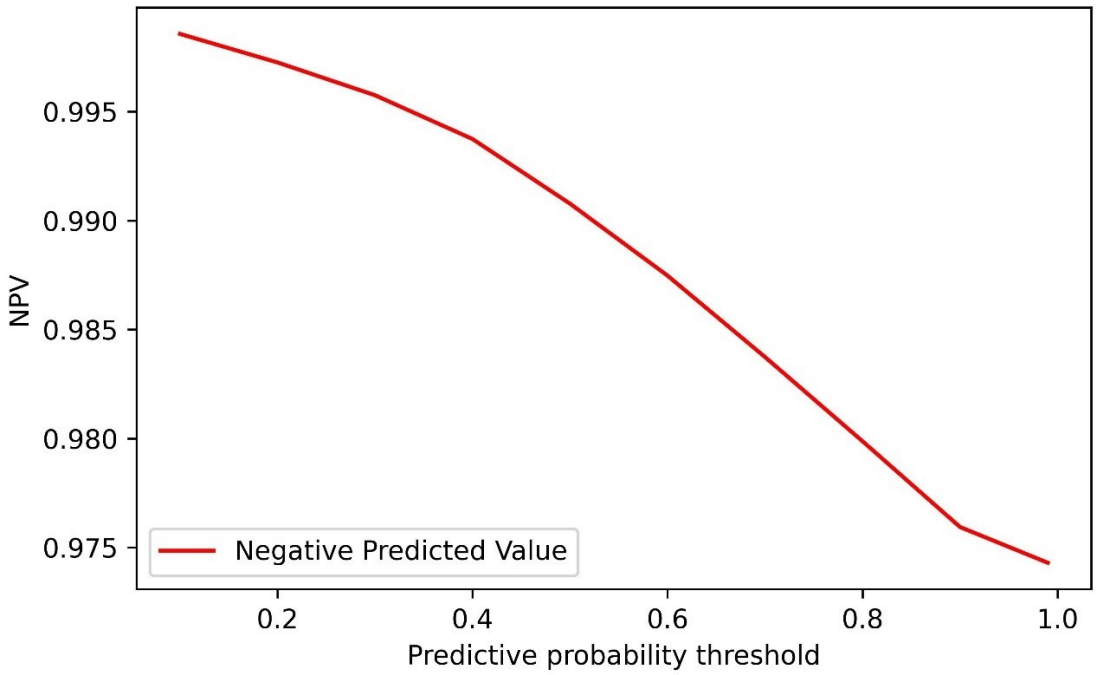
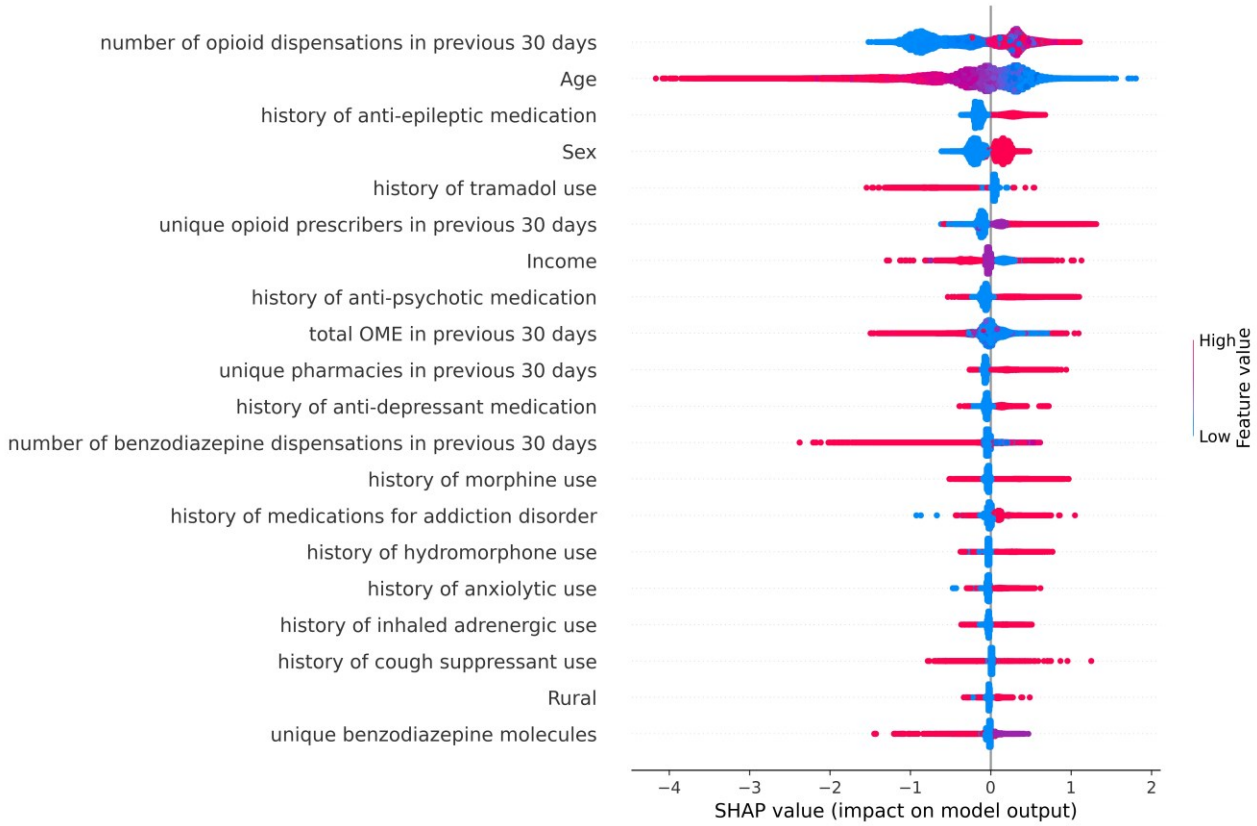


Figure 6.10. Negative predicted value vs predicted probability using the 2019 validation set.



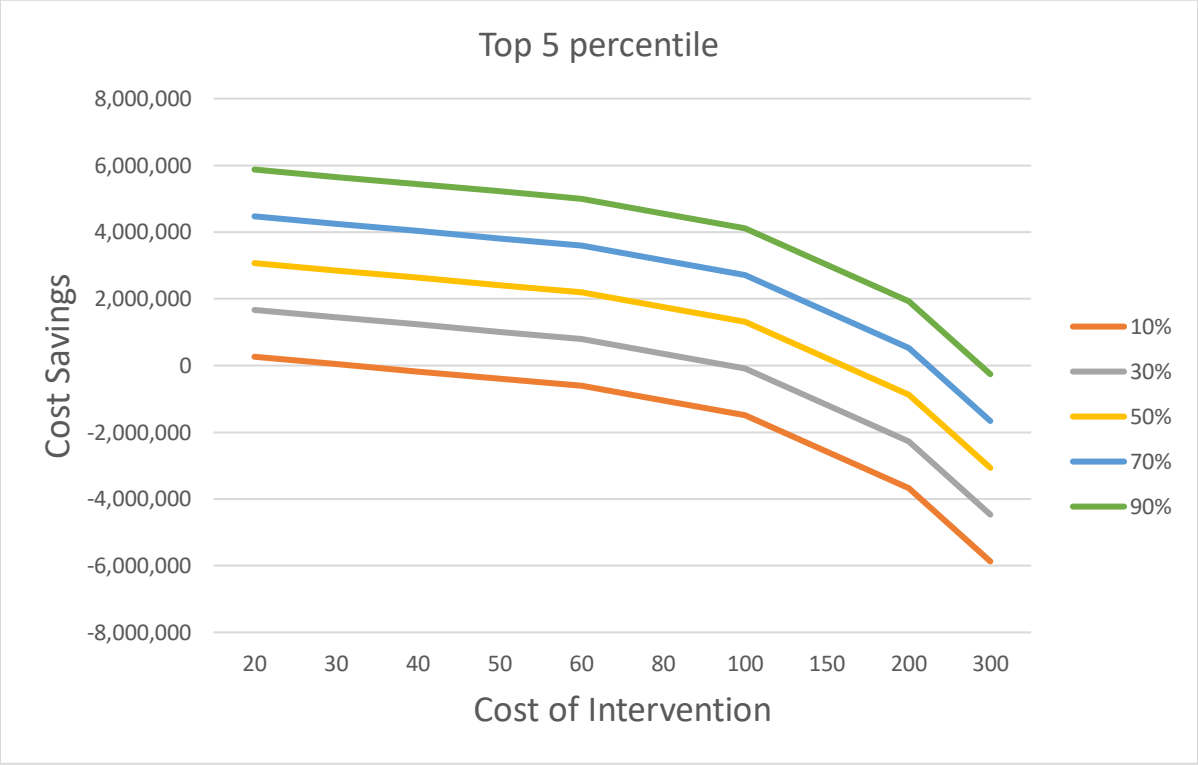
NPV: negative predicted value

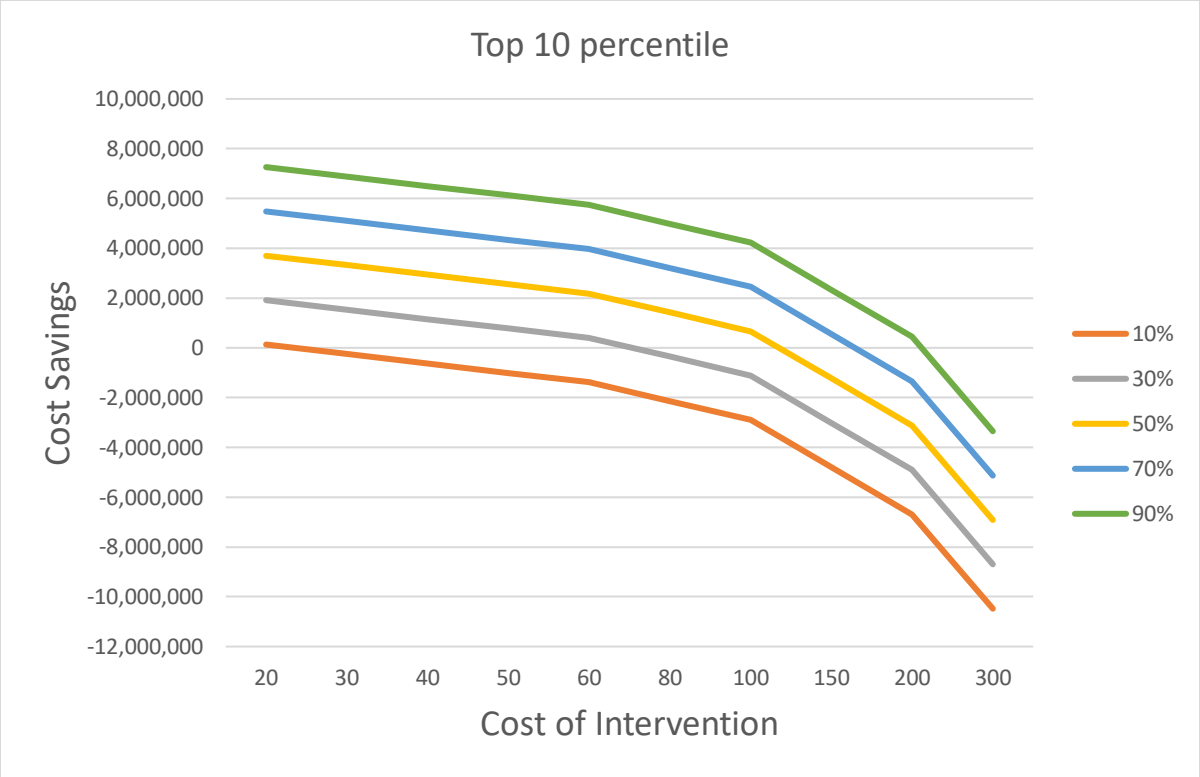
Figure 6.11. SHAP values and feature impact from the XGBoost classifier using the 2019 validation set to describe variable importance in relation to the outcome. Features are arranged from highest to lowest impact on prediction.



These predictors are rank-ordered according to feature importance with importance decreasing from top to bottom; variable impact does not have a causal nor statistical meaning but is simply a measure of the influence a predictor has on the ML model output. Red indicates higher values of the predictors while blue indicates lower values and lines to the right of the 0.0 on the x-axis are suggestive of higher risk of our outcome while those to the left are less suggestive. For example, a higher number of previous opioid dispensations (red) are predictive of a 30-day event (red is to the right of the 0.0 on x-axis) where in some instances, a higher number of previous dispensations (red) has a higher influence on 30-day events (further to the right of the 0.0 on the x-axis). Similarly, for binary variables, history of antidepressant use (red colour and to the right of the 0.0 on the x-axis) is suggestive of 30-day events to varying extents.

Figure 6.12. Cost savings and cost of interventions stratified by intervention success rate for 2019 Quarter 2 reported for both the top 5 and 10 percentiles of predicted risk. All costs are in Canadian dollars.




Top 10 percentile
10,000,000
8,000,000
6,000,000
4,000,000
2,000,000
0
-2,000,000
-4,000,000
-6,000,000
-8,000,000
-10,000,000
-12,000,000
Cost Savings
20
30
40
50
60
80
100
150
200
300
Cost of Intervention
10%
30%
50%
70%
90%

References for Chapter 6

6. O'Connor S, Grywacheski V, Louie K. At-a-glance - Hospitalizations and emergency department visits due to opioid poisoning in Canada. *Health Promotion and Chronic Disease Prevention in Canada.* 2018;38(6):244-247.
7. 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.
11. 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. Alberta Machine Intelligence Institute. Machine Learning Process Lifecycle. In:2019.
16. Morgenstern JD, Buajitti E, O'Neill M, et al. Predicting population health with machine learning: a scoping review. *BMJ Open.* 2020;10(10):e037860.
17. 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.
18. Luo W, Phung D, Tran T, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res.* 2016;18(12):e323.
19. 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.
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. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *Bmj.* 2016;352:i6.
23. 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.
24. 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.
34. Mooney SJ, Pejaver V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annual Review of Public Health.* 2018;39(1):95-112.
37. 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.

38. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European heart journal.* 2014;35(29):1925-1931.

43. Ravaut M, Harish V, Sadeghi H, et al. Development and validation of a machine learning model using administrative health data to predict onset of type 2 diabetes. *JAMA network open.* 2021;4(5):e2111315-e2111315.

50. 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.

57. Canadian Institute for Health Information. Patient Cost Estimator: Methodology Notes and Glossary. 2022; https://www.cihi.ca/sites/default/files/document/patient-cost-estimator-methodology-notes-2021-en.pdf. Accessed May 2022, 2022.

58. Glussich A. Estimating Costs of Hospital Stays. 2016 CADTH Symposium; 2016; Ottawa, ON, Canada.

59. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA.* 2018;320(21):2199-2200.

66. 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.

68. 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.

71. Reznick R HK, Horsely T, Hassani MS. *Task Force Report on Artificial Intelligence and Emerging Digital Technologies.* 2020.

73. Dowell D. CDC guideline for prescribing opioids for chronic pain. 2016.

77. World health Organization. Classification of Diseases (ICD). 2019; https://www.who.int/classifications/icd/icdonlineversions/en/. Accessed Jun 2020.

80. World Health Organization. International language for drug utilization research, ATC/DDD. 2020; https://www.whocc.no/. Accessed Jun 2020, 2020.

82. 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.

83. Brownlee J. A Gentle Introduction to Imbalanced Classification. 2020; https://machinelearningmastery.com/what-is-imbalanced-classification/. Accessed Jan 2021.

84. King G, Zeng L. Logistic regression in rare events data. *Political analysis.* 2001;9(2):137-163.

86. 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.

88. 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.

89. 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.

91. Molnar C. *Interpretable machine learning. A Guide for Making Black Box Models Explainable.* 2019.

92. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Paper presented at: Advances in neural information processing systems2017.

94. 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.

95. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Paper presented at: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining2016.

96. The pandas development team. pandas-dev/pandas: Pandas. 2020; https://doi.org/10.5281/zenodo.3509134, Jan 2021.

109. Alberta College of Pharmacy. 2019; https://abpharmacy.ca/. Accessed Sept 2019.

110. Canadian Institute for Health Information. 2019; https://www.cihi.ca/en.

120. XGBoost. Python API Reference. https://xgboost.readthedocs.io/en/latest/python/python_api.html#module-xgboost.sklearn. Accessed August 2021.

125. Ashfaq A, Sant'Anna A, Lingman M, Nowaczyk S. Readmission prediction using deep learning on electronic health records. *Journal of Biomedical Informatics.* 2019;97:103256.

141. College of Physicians and Surgeons of Alberta. Tracked Prescription Program. 2021; https://www.tppalberta.ca/.

145. Jaeschke R, Guyatt GH, Sackett DL, Group EBMW. How to use an article about a diagnostic test. *Jama.* 1994;271(5):389-391.

146. Guan Q, Khuu W, Martins D, et al. Evaluating the early impacts of delisting high-strength opioids on patterns of prescribing in Ontario. *Health Promotion and Chronic Disease Prevention in Canada.* 2018;38(6):256-262.

147. Peterson ED. Machine Learning, Predictive Analytics, and Clinical Practice: Can the Past Inform the Present? *JAMA.* 2019;322(23):2283-2284.

148. Sharma V, Kulkarni V, Eurich DT, Kumar L, Samanani S. Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada. *BMJ Open.* 2021;11(5):e043964.

149. Pink GH, Bolley HB. Physicians in health care management: 3. Case Mix Groups and Resource Intensity Weights: an overview for physicians. *CMAJ: Canadian Medical Association Journal.* 1994;150(6):889.

150. Pink GH, Bolley HB. Physicians in health care management: 4. Case Mix Groups and Resource Intensity Weights: physicians and hospital funding. *CMAJ: Canadian Medical Association Journal.* 1994;150(8):1255.

151. Canadian Institute for Health Information. Your Health System. 2022; https://yourhealthsystem.cihi.ca/hsp/indepth?lang=en#/indicator/015/2/C20018/. Accessed May 2022, 2022.

152. Chu F OA, Kaul P. Canadian Case Mixed Groups (CMG+) Costing Proxy for Acute Myocardial Infarction. *Journal of Health & Medical Economics.* 2016;2(2:8):1-4.

153. Chu F, Ohinmaa A, Jacobs P, Zheng Y, Kaul P. Comparing Actual Patient Level Hospital Costs To The Canadian Cmg+ Costing Estimates For Acute Myocardial Infarction. *Value in Health.* 2014;17(7):A481.

Chapter 7: Discussion

This PhD project developed and evaluated ML classifiers which predicted adverse outcomes important to health system planners in at-risk populations at the individual level. The ML classifiers were assessed using a range of metrics not commonly seen in ML prediction literature studying population health outcomes. This included measures such as discrimination, calibration, pre and post-test probabilities, ML explainability, real-world simulations and potential cost savings, all considered informative by health system planners. This body of work suggests that ML models could provide health systems planners with useful information when interventions are based on ranked predictions as opposed to those based on absolute probability thresholds, a finding verified by potential cost savings analyses.

From the first study, the findings revealed that predictions from ML classifiers were more informative than guideline-based rules for predicting 30-day risks from opioids. The c-statistics (i.e., discrimination performance) for the ML models were estimated around 0.87 while the guideline-based rules were around 0.5, no better than chance alone. Furthermore, the top 5 percentile of ML predicted risks translated into higher post-test probabilities (around 14%) compared to the pre-test probability of 1.6%. Limitations mainly focused on data issues in which only administrative health data was available to construct ML models. Health systems may find ML predictions in this context informative for opioid stewardship programs.

In the second project, predicting readmissions in people with heart failure, even the best performing ML models (c-statistic= 0.65) trained on administrative health data did not provide substantially informative prediction performance. Indeed, they only generated moderate shifts from pre to post-test probabilities (a shift from 21 to 24%). The ML models performed slightly better than the LACE tool (length of stay, acuity of admission, comorbidities, emergency department visits) developed using logistic regression (c-statistic= 0.65 vs. 0.57). The reasons why the ML model did not provide substantially informative prediction performance, especially when performance was high with opioids in the first study, is unknown. Part of the performance issues with predictions in patients with heart failure may simply be related to the lack of additional important health information in this population which is not

available in administrative data (e.g., patients' daily weights for fluid monitoring, salt and other diet considerations known to induce decompensation) or to the more seemingly randomness of decompensation in patients with heart failure. Health systems should take note of these findings since readmissions are an important outcome driving healthcare costs and poor health outcomes and other types of data must be included in ML model development.

The notion that administrative data may not be sufficient is also supported by the findings from the third project. Unlike the opioid study, developing ML models using only administrative health data may not provide health systems with sufficient informative predictions to use as decision aids for potential interventions, especially if considering daily or quarterly classifications of benzodiazepine risks in older adults. C-statistic for the ML model was 0.75. Although this indicates strong discrimination performance, likelihood ratios from daily and quarterly classifications were less than 10 implying uninformative ML predictions that would only provide modest additional information to health systems. Again, developing ML models solely on administrative data is a limiting factor.

The fourth project's findings showed that a ML classifier developed specifically for a health regulator (College of Physicians and Surgeons of Alberta) could assist their opioid stewardship program and provide value to health systems even when developed on very limited administrative health data. Based on validation analysis, the ML model had a c-statistic of 0.82 with the highest categories of predicted risk reporting likelihood ratios around 28. This translated to a post-test probability of 43.1% from a pre-test probability of 2.6%. Further, simulated interventions on these high-risk categories realized potential costs savings to the health system depending on intervention success rates. In contrast, net benefit analysis findings revealed that intervening on absolute probability thresholds was not informative at any cut point. Having access to more types of training data could improve prediction performance and further increase the value of ML assisted opioid stewardship programs.

Overall, in the opioid and benzodiazepine projects, the ML classifiers provided informative predictions, especially in the higher categories of ranked predictions with higher likelihood ratios. The frequency of ML classifications also affects informativeness. In the

benzodiazepine project, ML predictions classified daily or weekly were not informative while those done yearly did provide some measure of utility. It is up to health system planners to decide if yearly classifications of benzodiazepine risk in older adults is useful. Further, the opioid risk ML classifiers performed better than guideline-based rules. Net benefit analysis reported that ML prediction in these settings may not provide health systems with informative predictions with interventions based on absolute thresholds. This may be related to the limitations of developing ML models solely on administrative data, as many factors are absent. The ML model predicting readmissions in people with heart failure did not produce informative predictions beyond pre-test baseline risk, a finding which health system planners should note, as readmissions related to heart failure are an important outcome. In fact, the simple risk calculator based on logistic regression models (LACE score) did not provide any predictive information either in the heart failure population. Again, this can partly be explained by data issues in which administration health data alone cannot describe the variation in outcome distribution in a highly heterogeneous population. Based on these findings, health system planners may find the opioid ML classifiers informative while the benzodiazepine and heart failure readmission ML models not so much. Indeed, health systems could realize potential cost savings by implementing ML assisted opioid stewardship programs. Improvements in data collection may change the utility of using ML in other patient populations like those using benzodiazepines and experiencing readmissions.

Focusing on ML and opioids, ML deployment into real-world settings to assist opioid stewardship programs will have some of the key issues identified in Chapter 2 despite informative prediction results in studies. There is an opportunity for ML to integrate into health systems to mitigate opioid prescribing risks. Health jurisdictions and regulators could regularly risk stratify and intervene on high-risk opioid instances within a population to reduce system costs. In this case, ML, by identifying risks at the individual level, provides a patient focused approach which is a mandatory component of integrated health systems[154]. Opportunities also exist for ML assisted opioid stewardship at the provider and patient level. Having ML predictions available at clinical encounters may induce behaviour changes such as modified

prescribing and monitoring. Of course, the effects of these opportunities still must be substantiated and properly evaluated in controlled studies.

However, deploying a ML opioid risk classifier also faces hindrances. Issues related to data are substantial barriers and include types of data for training, where the data comes from, validity and reliability of that data, and subsequent generalizability of the ML model. Privacy and consent to collect 'big data' also come more into play which will further add to the complexity of using all these data sources as the public becomes more aware of what their data is potentially being used for. Targeted information campaigns to keep the public as informed as possible will be required for any of these initiatives to be successful. Opioid users are a heterogeneous population comprised of a large contingent of marginalized individuals. Bias, transparency and interpretability of ML predictions are important considerations at the system, provider and patient levels. It is mandatory for a health system program such as an opioid stewardship program to explain why certain patients are flagged as high risk or not. Whether it be through bias or some other mechanism, interventions conducted by ML assisted programs must be monitored and held accountable so all can benefit. The final potential barrier affects all aspects of the system, that being trained personnel. As ML implementation becomes more mainstream, universities and other institutes will have to be nimble to pivot to provide the necessary precision health training required to implement, support, and evaluate ML programs. This is certainly lacking across most of Canada. Indeed, training programs are often very siloed with individuals trained in ML methods but have limited training on health or health systems and vice versa. New integrated, interdisciplinary programs will be required to fully train and harness the data that is available for ML prediction in health systems.

This PhD program addressed some of the key issues in ML previously discussed in Chapter 2. Although data issues remained (ML classifiers were limited to training on administrative data only), the analysis included comprehensive metrics and measures to evaluate ML classifiers that health system planners would find informative; this contrasts with many studies in the literature[16,17]. These metrics could also inform writers who are developing reporting guidelines. The issue of ML interpretability was addressed by including measures of variable importance and how different features influenced predictions. All of these will assist in

making ML more transparent to users to help facilitate uptake as opposed to “black box” technology which is unacceptable to health systems.

Future Research

This PhD program did not address other key topics in ML prediction that health systems must acknowledge and are areas for future research. Research topics related to bias in data, ML modeling and implementation were not included and need to be investigated. As well, this project did not touch on the areas of data governance, consent to collect data and privacy. Health system planners should note that all ML projects would benefit by including these pieces in their intervention programs. More work in these areas is required if ML is going to assist public health initiatives and become more mainstream. Other issues were also raised that should direct future work in ML prediction. These include: handling missing and correlated data, more research into ML informed intervention success rates and costs, and attributable benefit of ML prediction. Health systems will have to identify priority areas related to patient stratification, cost of ML assisted intervention programs, workload, monitoring ML performance, bias, and transparency, just to name a few.

Future research specific to data issues is strongly needed if ML is to be successfully integrated into health systems. The common theme throughout this PhD has been to develop ML models with only administrative health data mostly because it is readily available and structured. Relying solely on this type of data may be limiting the capabilities of ML assisted intervention programs. Data of interest exists that is not part of the administrative data sets but with private sector entities and people. Indeed, health systems could benefit by accessing other types of data collected from personal monitoring devices, effluent sources (e.g., online search terms) and other supplies of unstructured data. Research on how these additional sources of data incrementally benefit ML predictions and subsequent outcomes must also be conducted. Failure to not quickly incorporate other data into ML may result in the perception that ML is not useful for health predictions. This belief or attitude among health administrators may be difficult to change and further hinder the use of ML models in practice. Indeed, this has already occurred in the AI/ML world as 20 years ago AI was touted as the new future; yet, lack of computing resources, data, and other factors resulted in a ‘false start’ which has now only

been addressed twenty years later. The same could occur in the health system if ML models are not quickly shown to be beneficial to patients, health providers, and the system overall.

Another major area of future research revolves around training personnel within health systems on how to implement and use ML. Education programs must be developed to bridge the divides between ML, public health and health systems planning. The need for these programs is acknowledged by health regulators and systems alike. Health providers also need education about ML and its applications. Further, research into how best to engage health providers to use ML is needed as well as research on the impact of "alert fatigue" that health providers face daily; adding ML to an already busy EMR platform will contribute to more automated flags and perhaps oversights.

Developing and validating a ML classifier is a small step in the overall lifecycle of ML deployment into real world scenarios. Beyond research and academics, ML implementation may fall into a framework like any other medical device or pharmaceutical (see Figure 7.1). After validation, ML classifiers could still have a long and arduous journey before health systems can fully embrace them[29,31].

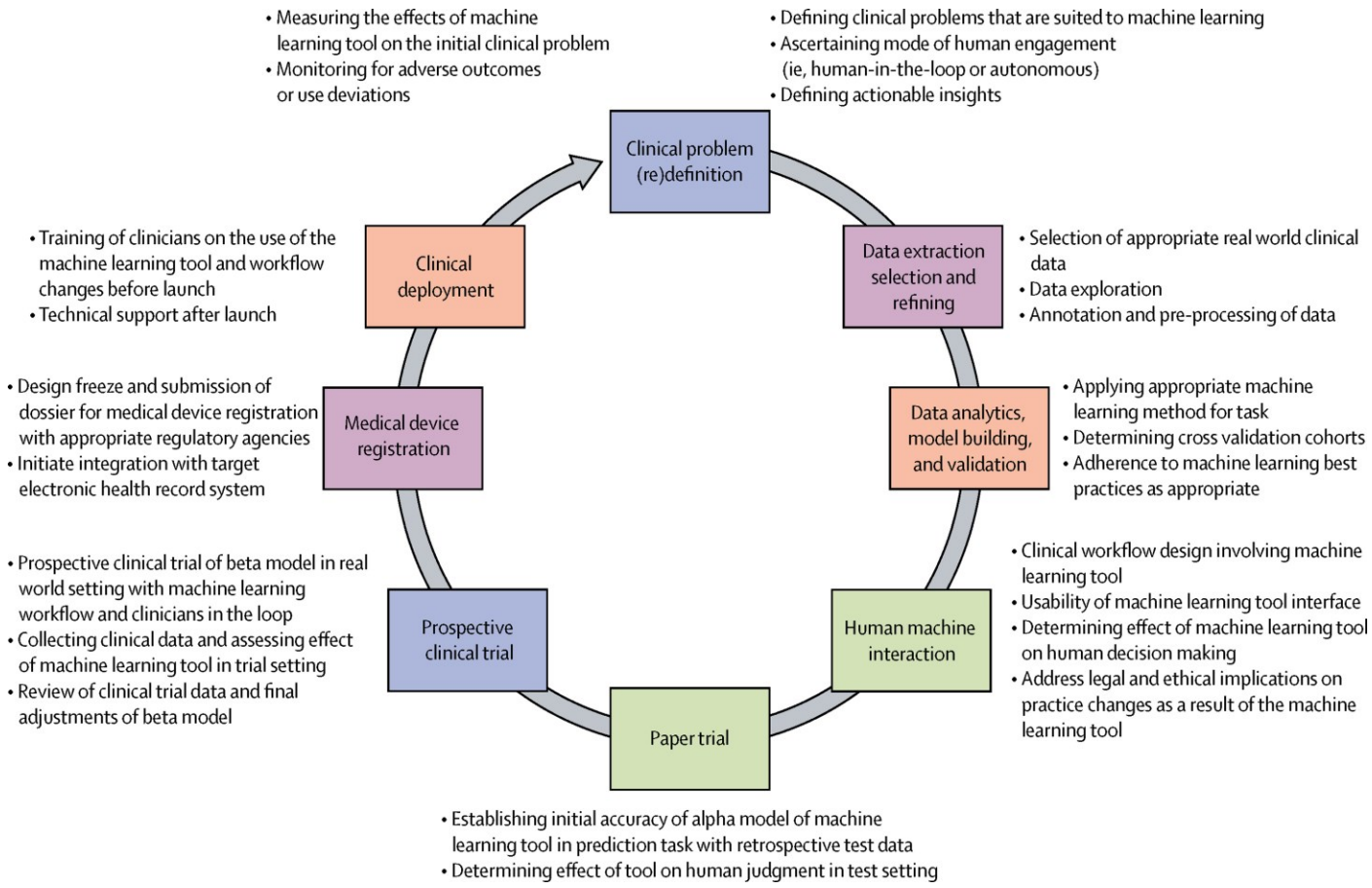


Figure 7.1. Implementation of ML into health-care applications. From *Lancet Oncol* 2019; 20: e262-273[29].

There are a lot of moving parts to account for when considering ML implementation into health systems and how best to evaluate them is still uncertain for some. Whether ML is categorized as a medical device requiring an intensive approval process or as an extension of information management remains unclear. Nevertheless, issues with data are at the core of these considerations. Collection of structured and unstructured data in a timely manner is important for real-time ML predictions. Storing this data in private infrastructure is necessary but in a way that allows for seamless sharing across data siloes. All these factors would have to be managed under a data governance framework. Health system planners, providers and the public will have to be educated on the use and benefits of ML prediction to counter the "black box" image. Further, the public will need convincing to share personal data that is not collected in administrative health data. Platforms like *MyHealth,* supported by the Government of

Alberta, could be used to upload and house additional types of data like dietary, exercise and other lifestyle habits. Finally, monitoring programs to ensure ML transparency need to be defined and established to ensure fairness, a key mandate for public health systems.

Conclusion

Ultimately, ML technology will integrate into health systems[71]. However, it is important to manage expectations because much hype surrounds the use of ML in health systems planning and the promise of this emerging technology needs to be eased against implementation challenges[29]; frameworks are needed to direct data governance and ethics. The role of ML within health systems is still unclear although the findings from this PhD program suggest its use in population health risk-based strategies is becoming increasingly evident and opportunities are emerging where this technology can help Albertans and others. However, it is important to understand that, as shown within this PhD program, ML methods are not a panacea. Although ML-based risk prediction strategies may work well enough in opioid stewardship, its impact was less informative in benzodiazepine risk prediction in older adults and of limited benefit in people with heart failure which is highly prevalent and costly to health systems. Indeed, rigorous evaluation of any ML-based program is necessary as successful deployment of ML assisted public health strategies will likely be specific to each disease state, data availability, population and even health system.

References for Discussion

16. Morgenstern JD, Buajitti E, O'Neill M, et al. Predicting population health with machine learning: a scoping review. *BMJ Open.* 2020;10(10):e037860.

17. 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.

29. Ngiam KY, Khor W. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology.* 2019;20(5):e262-e273.

31. Komorowski M. Clinical management of sepsis can be improved by artificial intelligence: yes. *Intensive Care Medicine.* 2020;46(2):375-377.

71. Reznick R HK, Horsely T, Hassani MS. *Task Force Report on Artificial Intelligence and Emerging Digital Technologies.* 2020.

154. Suter E, Oelke ND, Adair CE, Armitage GD. Ten key principles for successful health systems integration. *Healthcare quarterly (Toronto, Ont).* 2009;13 Spec No(Spec No):16-23.

Bibliography

1. Chechulin Y, Nazerian A, Rais S, Malikov K. Predicting patients with high risk of becoming high-cost healthcare users in Ontario (Canada). *Healthcare Policy.* 2014;9(3):68.
2. Curtis LJ, MacMinn WJ. Health Care Utilization in Canada: Twenty-five Years of Evidence. *Canadian Public Policy / Analyse de Politiques.* 2008;34(1):65-87.
3. Ploeg J, Matthew-Maich N, Fraser K, et al. Managing multiple chronic conditions in the community: a Canadian qualitative study of the experiences of older adults, family caregivers and healthcare providers. *BMC Geriatrics.* 2017;17(1):40.
4. Jencks SF, Williams MV, Coleman EA. Rehospitalizations among Patients in the Medicare Fee-for-Service Program. *New England Journal of Medicine.* 2009;360(14):1418-1428.
5. Alberta Health Services. *AHS Report on Performance FY 2017-18. Unplanned Medical Readmissions.* 2018.
6. O'Connor S, Grywacheski V, Louie K. At-a-glance - Hospitalizations and emergency department visits due to opioid poisoning in Canada. *Health Promotion and Chronic Disease Prevention in Canada.* 2018;38(6):244-247.
7. 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.
8. Orpana HM, Lang JJ, Baxi M, et al. Canadian trends in opioid-related mortality and disability from opioid use disorder from 1990 to 2014 through the lens of the Global Burden of Disease Study. *Health Promot Chronic Dis Prev Can.* 2018;38(6):234-243.
9. Canadian Institute for Health Information. All-Cause Readmission to Acute Care and Return to the Emergency Department. 2012.
10. Gawande A. The hot spotters. *The New Yorker.* 2011;86(45):40-51.
11. 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.
12. Bastanlar Y, Ozuysal M. Introduction to machine learning. *Methods in molecular biology (Clifton, NJ).* 2014;1107:105-128.
13. 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.
14. Alberta Machine Intelligence Institute. Machine Learning Process Lifecycle. In:2019.

15. Shah NH, Milstein A, Bagley P, Steven C. Making Machine Learning Models Clinically Useful. *JAMA.* 2019;322(14):1351-1352.
16. Morgenstern JD, Buajitti E, O'Neill M, et al. Predicting population health with machine learning: a scoping review. *BMJ Open.* 2020;10(10):e037860.
17. 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.
18. Luo W, Phung D, Tran T, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res.* 2016;18(12):e323.
19. 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.
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. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *Bmj.* 2016;352:i6.
22. 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.
23. 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.
24. 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.
25. Frankl SE, Breeling JL, Goldman L. Preventability of emergent hospital readmission. *The American journal of medicine.* 1991;90(6):667-674.
26. Cunningham CM, Hanley GE, Morgan S. Patterns in the use of benzodiazepines in British Columbia: examining the impact of increasing research and guideline cautions against long-term use. *Health Policy.* 2010;97(2):122-129.
27. Weir D. Benzodiazepine Receptor Agonist & Z-Drug Dispensations in Alberta: A Population - Based Descriptive Study2015.

28. ChooseWiselyCanada. The Canadian Geriatrics Society has developed a list of 5 things physicians and patients should question in geriatrics [Internet]. https://choosingwiselycanada.org/geriatrics/.
29. Ngiam KY, Khor W. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology.* 2019;20(5):e262-e273.
30. sonix.ai. What's the difference between artificial intelligence (AI), machine learning (ML) and natural language processing (NLP)? https://sonix.ai/articles/difference-between-artificial-intelligence-machine-learning-and-natural-language-processing. Accessed August 2022.
31. Komorowski M. Clinical management of sepsis can be improved by artificial intelligence: yes. *Intensive Care Medicine.* 2020;46(2):375-377.
32. Cognilytica. Cognitive Project Management for Artifical Intelligence Methodology. In:2020.
33. AI Enabled Care. Paper presented at: Building Collaboration for Deeper Learning and Better Care; April 29, 2021, 2021; Virtual.
34. Mooney SJ, Pejaver V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annual Review of Public Health.* 2018;39(1):95-112.
35. Canadian Institutes of Health Research. The Knowledge to Action Process. 2016; https://cihr-irsc.gc.ca/e/29418.html#6. Accessed August 2022.
36. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *JAMA.* 2018;319(13):1317-1318.
37. 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.
38. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European heart journal.* 2014;35(29):1925-1931.
39. Rose S. Machine Learning for Prediction in Electronic Health Data. *JAMA Network Open.* 2018;1(4):e181404-e181404.
40. Steyerberg EW. *Clincial Prediction Models.* Springer; 2009.
41. Michard F, Teboul JL. Predictive analytics: beyond the buzz. *Annals of Intensive Care.* 2019;9(1):46.
42. Frizzell JD, Liang L, Schulte PJ, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA cardiology.* 2017;2(2):204-209.

43. Ravaut M, Harish V, Sadeghi H, et al. Development and validation of a machine learning model using administrative health data to predict onset of type 2 diabetes. *JAMA network open.* 2021;4(5):e2111315-e2111315.
44. Xie H, McHugo G, Drake R, Sengupta A. Using discrete-time survival analysis to examine patterns of remission from substance use disorder among persons with severe mental illness. *Mental Health Services Research.* 2003;5(1):55-64.
45. Donders ART, Van Der Heijden GJ, Stijnen T, Moons KG. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology.* 2006;59(10):1087-1091.
46. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. *Journal of Clinical Epidemiology.*
47. Bhaskaran K, Smeeth L. What is the difference between missing completely at random and missing at random? *International Journal of Epidemiology.* 2014;43(4):1336-1339.
48. Greenland S, Finkle WD. A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *American Journal of Epidemiology.* 1995;142(12):1255-1264.
49. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research.* 2011;20(1):40-49.
50. 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.
51. Hu Z, Melton GB, Arsoniadis EG, Wang Y, Kwaan MR, Simon GJ. Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *Journal of Biomedical Informatics.* 2017;68:112-120.
52. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological methods.* 2002;7(2):147.
53. Sharafoddini A, Dubin JA, Maslove DM, Lee J. A new insight into missing data in intensive care unit patient profiles: Observational study. *JMIR medical informatics.* 2019;7(1):e11605.
54. Fletcher Mercaldo S, Blume JD. Missing data and prediction: the pattern submodel. *Biostatistics.* 2020;21(2):236-252.
55. Ding Y, Simonoff JS. An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research.* 2010;11(1).

56. Talhouk Aline. AI Predictive Analytics: pathways from research to the clinic. Paper presented at: AI Enabled Care: Building Collaboration for Deeper Learning and Better Care; April 2021, 2021; Michener Institute.
57. Canadian Institute for Health Information. Patient Cost Estimator: Methodology Notes and Glossary. 2022; https://www.cihi.ca/sites/default/files/document/patient-cost-estimator-methodology-notes-2021-en.pdf. Accessed May 2022, 2022.
58. Glussich A. Estimating Costs of Hospital Stays. 2016 CADTH Symposium; 2016; Ottawa, ON, Canada.
59. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA.* 2018;320(21):2199-2200.
60. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Paper presented at: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining2015.
61. McGinn TG, Guyatt GH, Wyer PC, et al. Users' Guides to the Medical LiteratureXXII: How to Use Articles About Clinical Decision Rules. *JAMA.* 2000;284(1):79-84.
62. Karthikeyan G, W. Eikelboom J. The CHADS2 score for stroke risk stratification in atrial fibrillation – friend or foe? *Thromb Haemost.* 2010;104(07):45-48.
63. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine.* 2018;178(11):1544-1547.
64. Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. *JAMA.* 2019;322(24):2377-2378.
65. Richter AN, Khoshgoftaar TM. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artificial Intelligence in Medicine.* 2018;90:1-14.
66. 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.
67. Webster LR, Webster RM. Predicting Aberrant Behaviors in Opioid-Treated Patients: Preliminary Validation of the Opioid Risk Tool. *Pain Medicine.* 2005;6(6):432-442.
68. 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.

69. Brindle PM, McConnachie A, Upton MN, Hart CL, Smith GD, Watt GC. The accuracy of the Framingham risk-score in different socioeconomic groups: a prospective study. *British Journal of General Practice.* 2005;55(520):838-845.
70. Advances in AI and Genomics: Creating a Revolution in Healthcare. May 21, 2021, 2021; Edmonton, Alberta.
71. Reznick R HK, Horsely T, Hassani MS. *Task Force Report on Artificial Intelligence and Emerging Digital Technologies.* 2020.
72. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447-453.
73. Dowell D. CDC guideline for prescribing opioids for chronic pain. 2016.
74. ismp Canada. Essential Clinical Skills for Opioid Prescribers. 2017; https://www.ismp-canada.org/download/OpioidStewardship/Opioid-Prescribing-Skills.pdf. Accessed Nov 2018.
75. Centre for Effective Practice. Management of Chronic Non Cancer Pain. 2017; thewellhealth.ca/cncp.
76. College of Physicians and Surgeons of Alberta. TPP Alberta – OME and DDD Conversion Factors. 2020; http://www.cpsa.ca/tpp/. Accessed Jun 2020.
77. World health Organization. Classification of Diseases (ICD). 2019; https://www.who.int/classifications/icd/icdonlineversions/en/. Accessed Jun 2020.
78. Gomes T, Mamdani MM, Dhalla IA, Paterson JM, Juurlink DN. Opioid Dose and Drug-Related Mortality in Patients With Nonmalignant PainOpioid Dose and Drug-related Mortality. *JAMA Internal Medicine.* 2011;171(7):686-691.
79. 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.
80. World Health Organization. International language for drug utilization research, ATC/DDD. 2020; https://www.whocc.no/. Accessed Jun 2020, 2020.
81. 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.
82. 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.
83. Brownlee J. A Gentle Introduction to Imbalanced Classification. 2020; https://machinelearningmastery.com/what-is-imbalanced-classification/. Accessed Jan 2021.

84. King G, Zeng L. Logistic regression in rare events data. *Political analysis.* 2001;9(2):137-163.
85. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. *Journal of Big Data.* 2019;6(1):1-54.
86. 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.
87. 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.
88. 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.
89. 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.
90. Shah ND, Steyerberg EW, Kent DM. Big Data and Predictive Analytics: Recalibrating Expectations. *JAMA.* 2018;320(1):27-28.
91. Molnar C. *Interpretable machine learning. A Guide for Making Black Box Models Explainable.* 2019.
92. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Paper presented at: Advances in neural information processing systems2017.
93. 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.
94. 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.
95. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Paper presented at: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining2016.
96. The pandas development team. pandas-dev/pandas: Pandas. 2020; https://doi.org/10.5281/zenodo.3509134, Jan 2021.
97. Ezekowitz JA, O'Meara E, McDonald MA, et al. 2017 Comprehensive Update of the Canadian Cardiovascular Society Guidelines for the Management of Heart Failure. *Canadian Journal of Cardiology.* 2017;33(11):1342-1433.
98. Yam CH, Wong EL, Chan FW, et al. Avoidable readmission in Hong Kong-system, clinician, patient or social factor? *BMC health services research.* 2010;10(1):311.

99. Alberta Health. Performance Measure Definition-30 day Overall Readmission Rate. 2014; https://open.alberta.ca/dataset/c7e3fc16-7aea-455c-96a1-20811a640b1a/resource/63ee45db-a066-4298-b63d-ba254eee5dc5/download/PMD-30-Day-Readmission-Rate.pdf.
100. Feltner C, Jones CD, Cené CW, et al. Transitional Care Interventions to Prevent Readmissions for Persons With Heart Failure. *Annals of Internal Medicine.* 2014;160(11):774-784.
101. van Walraven C, Dhalla IA, Bell C, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Canadian Medical Association Journal.* 2010;182(6):551-557.
102. Au AG, McAlister FA, Bakal JA, Ezekowitz J, Kaul P, van Walraven C. Predicting the risk of unplanned readmission or death within 30 days of discharge after a heart failure hospitalization. *American Heart Journal.* 2012;164(3):365-372.
103. Wang H, Robinson RD, Johnson C, et al. Using the LACE index to predict hospital readmissions in congestive heart failure patients. *BMC Cardiovascular Disorders.* 2014;14(1):97.
104. Yazdan-Ashoori P, Lee SF, Ibrahim Q, Van Spall HG. Utility of the LACE index at the bedside in predicting 30-day readmission or death in patients hospitalized with heart failure. *American heart journal.* 2016;179:51-58.
105. Mortazavi BJ, Downing NS, Bucholz EM, et al. Analysis of machine learning techniques for heart failure readmissions. *Circulation: Cardiovascular Quality and Outcomes.* 2016;9(6):629-640.
106. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology.* 2019;110:12-22.
107. Shin S, Austin PC, Ross HJ, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC Heart Failure.* 2021;8(1):106-115.
108. Ghimire A, Fine N, Ezekowitz JA, Howlett J, Youngson E, McAlister FA. Frequency, predictors, and prognosis of ejection fraction improvement in heart failure: an echocardiogram-based registry study. *European heart journal.* 2019;40(26):2110-2117.
109. Alberta College of Pharmacy. 2019; https://abpharmacy.ca/. Accessed Sept 2019.
110. Canadian Institute for Health Information. 2019; https://www.cihi.ca/en.
111. van Walraven C, Wong J, Forster AJ. LACE+ index: extension of a validated index to predict early death or urgent readmission after hospital discharge using administrative data. *Open Medicine.* 2012;6(3):e80.
112. Keyko K JL. *Pathway Pearls LACE Index Scoring.* Alberta Health Services; June 2018 2018.

113. Jiang W, Siddiqui S, Barnes S, et al. Readmission Risk Trajectories for Patients With Heart Failure Using a Dynamic Prediction Approach: Retrospective Study. *JMIR medical informatics.* 2019;7(4):e14756-e14756.
114. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Chow BJ, Dwivedi G. Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death. *PLOS ONE.* 2019;14(6):e0218760.
115. Leppin AL, Gionfriddo MR, Kessler M, et al. Preventing 30-day hospital readmissions: a systematic review and meta-analysis of randomized trials. *JAMA internal medicine.* 2014;174(7):1095-1107.
116. Vader JM, LaRue SJ, Stevens SR, et al. Timing and Causes of Readmission After Acute Heart Failure Hospitalization-Insights From the Heart Failure Network Trials. *J Card Fail.* 2016;22(11):875-883.
117. Reddy YN, Borlaug BA. Readmissions in Heart failure: It's more than just the medicine. Paper presented at: Mayo Clinic Proceedings2019.
118. Retrum JH, Boggs J, Hersh A, et al. Patient-identified factors related to heart failure readmissions. *Circ Cardiovasc Qual Outcomes.* 2013;6(2):171-177.
119. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research.* 2011;12:2825-2830.
120. XGBoost. Python API Reference. https://xgboost.readthedocs.io/en/latest/python/python_api.html#module-xgboost.sklearn. Accessed August 2021.
121. Ali M. PyCaret: An open source, low-code machine learning library in Python version 2.3. April 2020; https://pycaret.org/about.
122. Jin H, Song Q, Hu X. Auto-keras: An efficient neural architecture search system. Paper presented at: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining2019.
123. H20.ai. 2021; https://www.h2o.ai/company/.
124. Schmidhuber J, Hochreiter S. Long short-term memory. *Neural Comput.* 1997;9(8):1735-1780.
125. Ashfaq A, Sant'Anna A, Lingman M, Nowaczyk S. Readmission prediction using deep learning on electronic health records. *Journal of Biomedical Informatics.* 2019;97:103256.

126. Dharmarajan K, Hsieh AF, Lin Z, et al. Diagnoses and Timing of 30-Day Readmissions After Hospitalization for Heart Failure, Acute Myocardial Infarction, or Pneumonia. *JAMA.* 2013;309(4):355-363.
127. Alberta Health Services and Government of Alberta. Admissions for Ambulatory Care Sensitive Conditions Indicator Definition. 2011.
128. van Smeden M, Groenwold RHH, Moons KGM. A cautionary note on the use of the missing indicator method for handling missing data in prediction research. *Journal of Clinical Epidemiology.*
129. Pencina MJ, D'Agostino RB, Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in medicine.* 2011;30(1):11-21.
130. Mitnitski A, Howlett SE, Rockwood K. Heterogeneity of Human Aging and Its Assessment. *The Journals of Gerontology: Series A.* 2016;72(7):877-884.
131. Liu L-F. The Health Heterogeneity of and Health Care Utilization by the Elderly in Taiwan. *International Journal of Environmental Research and Public Health.* 2014;11(2):1384-1397.
132. Bayati M, Braverman M, Gillam M, et al. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PloS one.* 2014;9(10):e109264.
133. Komorowski M. Artificial intelligence in intensive care: are we there yet? *Intensive care medicine.* 2019;45(9):1298-1300.
134. Alberta Health. Alberta Health Services performance review : summary report. 2020; https://open.alberta.ca/publications/alberta-health-services-performance-review-summary-report#summary.
135. Katzman MA, Bleau P, Blier P, et al. Canadian clinical practice guidelines for the management of anxiety, posttraumatic stress and obsessive-compulsive disorders. *BMC Psychiatry.* 2014;14 Suppl 1:S1.
136. Pottie K, Thompson W, Davies S, et al. Deprescribing benzodiazepine receptor agonists. 2018.
137. Canadian Pharmacists Association. RxTx. 2019; https://www.e-therapeutics.ca/search.
138. CPSA. Clinical Toolkit Benzodiazepines: Use and Taper. *CPSA.* 2015.
139. TOP TOP. Guideline for Adult Primary Insomnia [Internet]. 2010; http://www.topalbertadoctors.org/download/439/insomnia_management_guideline.pdf.
140. Weir DL, Samanani S, Gilani F, Jess E, Eurich DT. Benzodiazepine receptor agonist dispensations in Alberta: a population-based descriptive study. *CMAJ Open.* 2018;6(4):E678-E684.

141. College of Physicians and Surgeons of Alberta. Tracked Prescription Program. 2021; https://www.tppalberta.ca/.
142. American Geriatrics Society Beers Criteria Update Expert P. American Geriatrics Society updated Beers Criteria for potentially inappropriate medication use in older adults. *J Am Geriatr Soc.* 2012;60(4):616-631.
143. O'Mahony D. STOPP/START criteria for potentially inappropriate medications/potential prescribing omissions in older people: origin and progress. *Expert review of clinical pharmacology.* 2020;13(1):15-22.
144. Urquhart R, Giguere AM, Lawson B, et al. Rules to identify persons with frailty in administrative health databases. *Canadian Journal on Aging/La Revue canadienne du vieillissement.* 2017;36(4):514-521.
145. Jaeschke R, Guyatt GH, Sackett DL, Group EBMW. How to use an article about a diagnostic test. *Jama.* 1994;271(5):389-391.
146. Guan Q, Khuu W, Martins D, et al. Evaluating the early impacts of delisting high-strength opioids on patterns of prescribing in Ontario. *Health Promotion and Chronic Disease Prevention in Canada.* 2018;38(6):256-262.
147. Peterson ED. Machine Learning, Predictive Analytics, and Clinical Practice: Can the Past Inform the Present? *JAMA.* 2019;322(23):2283-2284.
148. Sharma V, Kulkarni V, Eurich DT, Kumar L, Samanani S. Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada. *BMJ Open.* 2021;11(5):e043964.
149. Pink GH, Bolley HB. Physicians in health care management: 3. Case Mix Groups and Resource Intensity Weights: an overview for physicians. *CMAJ: Canadian Medical Association Journal.* 1994;150(6):889.
150. Pink GH, Bolley HB. Physicians in health care management: 4. Case Mix Groups and Resource Intensity Weights: physicians and hospital funding. *CMAJ: Canadian Medical Association Journal.* 1994;150(8):1255.
151. Canadian Institute for Health Information. Your Health System. 2022; https://yourhealthsystem.cihi.ca/hsp/indepth?lang=en#/indicator/015/2/C20018/. Accessed May 2022, 2022.
152. Chu F OA, Kaul P. Canadian Case Mixed Groups (CMG+) Costing Proxy for Acute Myocardial Infarction. *Journal of Health & Medical Economics.* 2016;2(2:8):1-4.

153. Chu F, Ohinmaa A, Jacobs P, Zheng Y, Kaul P. Comparing Actual Patient Level Hospital Costs To The Canadian Cmg+ Costing Estimates For Acute Myocardial Infarction. *Value in Health.* 2014;17(7):A481.

154. Suter E, Oelke ND, Adair CE, Armitage GD. Ten key principles for successful health systems integration. *Healthcare quarterly (Toronto, Ont).* 2009;13 Spec No(Spec No):16-23.