A Machine Learning Approach to Overdose Risk Assessment

by

Man Yeung (Andy) Tai

B.Sc., University of Toronto (St. George), 2017

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in
THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Neuroscience)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

April 2024

© Man Yeung (Andy) Tai, 2024
The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

A Machine Learning Approach to Overdose Risk Assessment

submitted by Man Yeung (Andy) Tai in partial fulfilment of the requirements for

the degree of Doctor of Philosophy in Neuroscience

Examining Committee:

Reinhard Michael Krausz, Professor, Department of Psychiatry, University of British Columbia
Supervisor

Alireza Kazemi, Data Scientist, Department of Psychiatry, University of British Columbia
Supervisory Committee Member

David Kealy, Associate Professor, Department of Psychiatry, University of British Columbia
University Examiner

Skye Barbic, Associate Professor, Department of Occupational Science and Occupational Therapy, University of British Columbia
University Examiner

Additional Supervisory Committee Members:

Raymond Ng, Professor, Department of Computer Science, University of British Columbia
Supervisory Committee Member

Christian Schütz, Professor, Department of Psychiatry, University of British Columbia
Supervisory Committee Member
Abstract

This project explores the escalating drug toxicity crisis in British Columbia, spotlighting the role of machine learning within eHealth solutions to combat this pressing public health issue. The crisis, rooted in a shift from prescription opioids to potent synthetic alternatives like fentanyl, necessitates innovative intervention strategies. The study leverages extensive data from 2015 to 2019, aiming to develop predictive models for overdoses to improve healthcare responses.

The research begins with a systematic review and meta-analysis of machine learning models targeting opioid-related outcomes, demonstrating the predictive strengths of various algorithms in cohort studies. Results indicate these algorithms' effectiveness in forecasting opioid usage and overdose risks. Additionally, a review of clinical decision support systems in addiction and mental health care reveals their critical impact on enhancing diagnosis, treatment, and patient care, based on randomized controlled trials.

Central to the thesis is an exploratory data analysis utilizing the British Columbia Provincial Overdose Cohort. This involves rigorous data preparation, including wrangling, addressing missing data, and correcting class imbalances. The study assesses several machine learning models, including ensemble approaches like Random Forest and XGBoost, for their predictive accuracy regarding fatal and general overdoses. Despite challenges, these models demonstrate significant potential, with the best performers achieving over 90% accuracy in predicting general overdoses, though models for fatal overdoses showed lower efficacy.

The thesis concludes with an affirmation of machine learning's transformative potential in personalizing addiction psychiatry treatment through eHealth innovations such as the Risk Assessment and Management Platform (RAMP). This novel approach aims to individualize
treatment and prevention strategies, contributing to global efforts in mitigating the mental health ramifications of the drug toxicity crisis and transforming healthcare practices. Through the strategic implementation of machine learning, the study underscores a promising avenue for advancing healthcare solutions tailored to the intricacies of addiction and overdose risks, reflecting a significant stride towards mitigating the public health impacts of the drug crisis.
Lay Summary

This research project delves into the drug toxicity crisis in British Columbia, tracing its roots to systemic issues and evolving drug use patterns. The focus is on how data, collected from 2015-2019, can inform eHealth solutions like the Risk Assessment and Management Platform (RAMP) to address this crisis. Utilizing machine learning, RAMP aims to personalize treatment and prevention strategies.

The dissertation includes a systematic review of machine learning in predicting opioid-related outcomes, highlighting the effectiveness of specific models. It also examines clinical decision support systems in addiction and mental health care, assessing their impact through randomized trials.

The project underscores the potential of machine learning in predicting both fatal and general overdoses. Despite challenges in data collection and model limitations, the study shows the significant role of these technologies in advancing personalized medicine in addiction psychiatry, potentially transforming healthcare approaches to the drug toxicity crisis.
Preface

Under the mentorship of Professor Michael Krausz, I was introduced to the potential of utilizing electronic health platforms to contribute to resolving the drug toxicity crisis. My background in machine learning applications in psychiatry anchored my resolve to be a part of the solution in my hometown amidst the ongoing crisis in British Columbia. My connection to Downtown Vancouver, where I spent my childhood years, deepened my commitment. Returning in 2018 after my undergraduate studies at the University of Toronto, I was dismayed to witness the deterioration of the city. The drug toxicity crisis, I observed, had intensified.

With the invaluable support of the ACD group, including the insights of Dr. Alireza Kazemi and Dr. Kerry Jang, and under the vigilant supervision of Dr. Michael Krausz, I embarked on two systematic reviews. These reviews were pivotal in gathering evidence to support the use of machine learning for predicting outcomes in the overdose population. The next phase of my research was made possible by Dr. Amanda Slaunwhite of the British Columbia Centre for Disease Control, who facilitated access to the BC Provincial Overdose Cohort — the dataset essential for developing the predictive models.

The guidance of Professor Krausz, along with the expertise of Dr. Alireza Kazemi, Professor Christian Schutz, and Professor Raymond Ng, was instrumental in bringing this project to fruition. My role in this endeavor was to carry out the data analysis and interpretation, under the guidance of Professor Krausz. Initially I had access to the data through a remote desktop through Dr. Amanda Slaunwhite’s research lab from the BCCDC where I initially worked on the project. Subsequently, in 2022 I applied for data access through Population Data British Columbia and completed this project.
Chapter 2 is a revised version of the manuscript titled: Clinical Decision Support Systems in Addiction and Concurrent Disorders: A Systematic Review and Meta-Analysis.

I was the first author of this publication and conducted the review and the write-up of the manuscript under the supervision of Professor Krausz. Below is the detailed list concerning the contribution of each author to the manuscript.

Man Yeung (Andy) Tai: developing the protocol, electronic search, appraisal and writing the manuscripts.

Alireza KAZemi: developing the protocol and revising the manuscript.

Jane J. Kim: developing the protocol, appraisal, and revising the manuscript.

Jim. Schmeckenbecher: developing the methods for meta-analysis and reporting results and revising the manuscript.

Vanessa Kitchin: developing the protocol and importing search and revising the manuscript.

Janet Suen: appraisal of results and revising the manuscript.

Ryan Moro: appraisal of results and revising the manuscript.

Michael Krausz: developing the protocol, manuscript write-up and revision.

Chapter 3 is a revised version of the manuscript titled: Clinical Decision Support Systems in Addiction and Concurrent Disorders: A Systematic Review and Meta-Analysis.

I was the first author of this publication and conducted the review and the write-up of the manuscript under the supervision of Professor Krausz. Below is the detailed list concerning the contribution of each author to the manuscript.

Man Yeung (Andy) Tai: developing the protocol, electronic search, appraisal and writing the manuscripts.

Jane J. Kim: developing the protocol, appraisal, and revising the manuscript.
Jim. Schmeckenbecher: developing the methods for meta-analysis and reporting results and revising the manuscript.

Vanessa Kitchin: developing the protocol and importing search and revising the manuscript.

Johnston Wang: appraisal of results and revising the manuscript.

Alireza Kazemi: developing the protocol and revising the manuscript.

Frank Iorfino: revising the manuscript.

Michael Krausz: developing the protocol, manuscript write-up and revision.

Chapter 4 describes the initial exploratory data-analysis and is currently under review for publication. I will be the first author, with support from Alireza Kazemi and under the supervising of Professor Krausz.

Chapter 5 describes the development and evaluation of machine learning models predicting fatal overdose and is under review for publication. I will be the first author, with support from Alireza Kazemi and under the supervision of Professor Krausz.

Chapter 6 describes the development and evaluation of machine learning models predicting general overdose and is currently under review for publication. I will be the first author, with support from Alireza Kazemi and under the supervision of Professor Krausz.

Appendix A is the appendix for chapter 2.

Appendix B is the appendix for chapter 3.

Appendix C includes raw code utilized in creating the machine learning models predicting fatal and general overdose, from chapters 4-6.

The protocol for this study was reviewed and approved by UBC clinical Research ethics board (H20-02905).
# Table of Contents

Abstract ............................................................................................................................................... iii

Lay Summary ....................................................................................................................................... v

Preface ................................................................................................................................................ vi

Table of Contents ................................................................................................................................. ix

List of Tables ....................................................................................................................................... xviii

List of Figures ....................................................................................................................................... xx

List of Abbreviations ......................................................................................................................... xxiii

Glossary .............................................................................................................................................. xxiv

Acknowledgements ............................................................................................................................. xxvi

Dedication ............................................................................................................................................ xxviii

Chapter 1: Introduction ...................................................................................................................... 1

1.1 Institutional change and the rise of homelessness in Vancouver ........................................... 1

1.2 The emergence of opioids ......................................................................................................... 3

1.3 Current drug toxicity crisis ...................................................................................................... 6

1.4 Current healthcare approach .................................................................................................... 9

1.4.1 Overdose population ........................................................................................................... 13

1.4.1.1 Overdose prevention ....................................................................................................... 14

1.5 British Columbia Provincial Overdose Cohort ..................................................................... 15

1.5.1 An overview of the data ...................................................................................................... 16

1.5.2 Defining Cases and Controls in the Overdose Cohort ..................................................... 19

1.5.3 The integration of the BCCDC dataset for a E-health platform, a clinical decision support system ...................................................................................................................... 20
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6</td>
<td>Risk Assessment and Management Platform</td>
<td>21</td>
</tr>
<tr>
<td>1.6.1</td>
<td>E-Health Model</td>
<td>21</td>
</tr>
<tr>
<td>1.6.1.1</td>
<td>E-health in Canada</td>
<td>23</td>
</tr>
<tr>
<td>1.6.2</td>
<td>Risk Assessment and Management Platform</td>
<td>24</td>
</tr>
<tr>
<td>1.6.3</td>
<td>High Risk Assessment</td>
<td>27</td>
</tr>
<tr>
<td>1.6.3.1</td>
<td>Overdose risks</td>
<td>28</td>
</tr>
<tr>
<td>1.6.4</td>
<td>The Usage of the BC Provincial Overdose Cohort</td>
<td>29</td>
</tr>
<tr>
<td>1.7</td>
<td>Machine learning</td>
<td>30</td>
</tr>
<tr>
<td>1.7.1</td>
<td>Machine learning and statistics</td>
<td>32</td>
</tr>
<tr>
<td>1.7.2</td>
<td>Supervised machine learning problem</td>
<td>34</td>
</tr>
<tr>
<td>1.7.3</td>
<td>Machine learning in mental health</td>
<td>36</td>
</tr>
<tr>
<td>1.7.4</td>
<td>Project Aim</td>
<td>37</td>
</tr>
<tr>
<td>1.8</td>
<td>Research Question</td>
<td>38</td>
</tr>
<tr>
<td>1.9</td>
<td>An Overview of the following chapters</td>
<td>38</td>
</tr>
</tbody>
</table>

Chapter 2: Utilizing machine learning for early intervention and risk management in the opioid overdose crisis: a systematic review and meta-analysis in addiction psychiatry .......... 40

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>40</td>
</tr>
<tr>
<td>2.2</td>
<td>Methods</td>
<td>42</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Protocol</td>
<td>42</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Search Strategy</td>
<td>42</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Eligibility Criteria</td>
<td>43</td>
</tr>
<tr>
<td>2.2.4</td>
<td>Risk of bias</td>
<td>43</td>
</tr>
<tr>
<td>2.2.5</td>
<td>Data Extraction</td>
<td>43</td>
</tr>
</tbody>
</table>
Chapter 2: Methodology for Meta-Analysis

2.2.6 Methodology for Meta-Analysis ................................................................. 44

2.3 Results ........................................................................................................... 45

2.3.1 Meta-Analysis Results ............................................................................. 50

2.4 Discussion ..................................................................................................... 52

2.4.1 Popular models in predicting opioid related outcomes ....................... 53

2.4.2 Popular metrics in predicting opioid related outcomes ....................... 53

2.4.3 Data-Driven Insights .............................................................................. 54

2.4.4 Model performance comparing model types: Meta-Analysis Insights ...... 55

2.4.5 Model performance comparing model outcomes: Meta-Analysis Insights ..... 56

2.4.6 Future of Machine learning: Harnessing big data/Al and its limitations .... 57

2.5 Conclusion ................................................................................................... 59

2.5.1 Study Limitations ................................................................................... 59

2.6 Risk of Bias.................................................................................................... 60

Chapter 3: Clinical decision support systems in addictions and concurrent disorders: a
systematic review and meta analysis ................................................................. 64

3.1 Introduction .................................................................................................. 64

3.1.1 What is a Clinical Decision Support System? .................................... 65

3.2 Methods ......................................................................................................... 67

3.2.1 Protocol .................................................................................................. 67

3.2.2 Search Strategy ....................................................................................... 67

3.2.3 Eligibility Criteria .................................................................................. 68

3.2.4 Risk of Bias ............................................................................................ 68

3.2.5 Data Extraction ....................................................................................... 68
5.4.2 Machine learning results

5.5 Discussion

5.5.1 Model Performance Overview

5.5.2 Data and Model Challenges

Modeling for overdose: utilizing machine learning in predicting general overdose

5.6 Objective

5.7 Methods

5.7.1 Data Preprocessing

5.7.1.1 Sample size

5.7.1.2 Data splitting

5.7.1.3 Missing Data

5.7.1.4 Resampling

5.7.1.5 Machine learning

5.8 Results

5.8.1 Resampling results

5.8.2 Machine learning results

5.9 Discussion

5.9.1 Model Performance Overview

5.9.2 Comparative Analysis

5.10 Conclusion

Chapter 6: Exploratory secondary analysis, inference, and survival analysis

6.1 Objective

6.2 Stratifying risk factors in fatal overdose population
## 6.2.1 Methods

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.2.1.1 Fatal overdose dataset</td>
<td>143</td>
</tr>
<tr>
<td>6.2.1.2 Missing Data</td>
<td>144</td>
</tr>
</tbody>
</table>

## 6.2.2 Results

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.2.2.1 Variable importance</td>
<td>144</td>
</tr>
<tr>
<td>6.2.2.2 Logistic Regression</td>
<td>145</td>
</tr>
<tr>
<td>6.2.2.3 Correlational plot</td>
<td>149</td>
</tr>
</tbody>
</table>

## 6.2.3 Discussion

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.2.3.1 Chronic diseases</td>
<td>150</td>
</tr>
<tr>
<td>6.2.3.2 Substance use and mental illness</td>
<td>150</td>
</tr>
<tr>
<td>6.2.3.3 Figure interpretations</td>
<td>151</td>
</tr>
</tbody>
</table>

## 6.3 Stratifying risk factors in the general overdose

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3.1 Methods</td>
<td>153</td>
</tr>
<tr>
<td>6.3.1.1 Sample</td>
<td>153</td>
</tr>
<tr>
<td>6.3.1.2 Missing data</td>
<td>154</td>
</tr>
<tr>
<td>6.3.2 Results</td>
<td>154</td>
</tr>
<tr>
<td>6.3.2.1 Variable importance</td>
<td>154</td>
</tr>
<tr>
<td>6.3.2.2 Logistic Regression</td>
<td>155</td>
</tr>
<tr>
<td>6.3.2.3 Correlational plot</td>
<td>158</td>
</tr>
<tr>
<td>6.3.3 Discussion</td>
<td>158</td>
</tr>
<tr>
<td>6.3.3.1 Chronic diseases</td>
<td>159</td>
</tr>
<tr>
<td>6.3.3.2 Substance use and mental illness</td>
<td>159</td>
</tr>
<tr>
<td>6.3.3.3 Figure interpretations</td>
<td>160</td>
</tr>
</tbody>
</table>
6.4 Survival Analysis ............................................................................................................. 162

6.4.1 Methods ....................................................................................................................... 162

6.4.1.1 Sample ..................................................................................................................... 162

6.4.1.2 Packages used .......................................................................................................... 162

6.4.2 Results ......................................................................................................................... 163

6.4.3 Discussion .................................................................................................................... 166

6.4.4 Limitations .................................................................................................................. 167

Chapter 7: Conclusions ........................................................................................................ 169

Summary of Dissertation ...................................................................................................... 169

Summary of Introduction ....................................................................................................... 170

Summary of Chapter 2: Systematic review and meta-analysis on machine learning ........ 171

Summary of Chapter 3: Systematic review and meta-analysis on clinical decision support systems ...................................................................................................................................................... 173

Summary of Chapter 4: EDA and pilot modeling ................................................................. 174

Summary of Chapter 5: predicting fatal overdose ............................................................... 175

Summary of Chapter 5: Predicting general overdose .......................................................... 176

Summary of Chapter 6: Exploratory analysis and survival analysis .................................... 178

7.1 Significance ................................................................................................................... 179

7.1.1 Clinical Relevance ...................................................................................................... 179

7.1.2 Research Relevance .................................................................................................. 182

7.2 Limitations .................................................................................................................... 183

7.3 Future directions .......................................................................................................... 186

7.4 Concluding remarks ..................................................................................................... 187
Appendices

Appendix A Chapter 2 Appendix

A.1 Extraction Table
A.2 Search Strategies
A.3 All models and outcomes meta-analysis results

Appendix B Chapter 3

B.1 Search Strategies

Appendix C Chapter 4 Appendix

C.1 British Columbia Provincial Overdose Cohort Descriptions for Data Sources

Appendix D Computational Tools and Code

D.1 Packages utilized on RStudio
D.2 Code for forest plots
D.3 Data wrangling code
D.4 ML Prediction Code for Fatal Overdose
D.5 ML Prediction Code for Overdose
**List of Tables**

Table 1 Qualitative Analysis of studies for Addiction & Concur Disorders: Systematic Review
(Short, full version can be viewed in Appendix A) ................................................................. 46

Table 2 Risk of bias assessment for COHORT studies (196) for Machine Learning in Drug
toxicity crisis Management: A Systematic Review and Meta-Analysis ................................. 60

Table 3 Qualitative Analysis of studies for Machine Learning in Drug toxicity crisis
Management: A Systematic Review and Meta-Analysis (Papers highlighted with the same color
describe different phases of the same studies) ........................................................................... 72

Table 4 Risk of bias assessment for Randomized Control Trials, Risk Of Bias 2 (1) for Clinical
Decision Support in Addiction and Concurrent Disorders: A Systematic Review and Meta-
Analysis ...................................................................................................................................... 85

Table 5 Risk of bias assessment for COHORT studies, Modified Newcastle Ottawa Quality
Assessment Scale (196) for Clinical Decision Support in Addiction and Concurrent Disorders: A
Systematic Review and Meta-Analysis ....................................................................................... 86

Table 6 Risk of bias assessment for randomized controlled trials or case control studies,
Modified Newcastle Ottawa Quality Assessment Scale (196) for Clinical Decision Support in
Addiction and Concurrent Disorders: A Systematic Review and Meta-Analysis ..................... 87

Table 7 Risk of bias assessment for cross-sectional studies, Modified Newcastle Ottawa Quality
Assessment Scale (196) for Clinical Decision Support in Addiction and Concurrent Disorders: A
Systematic Review and Meta-Analysis ....................................................................................... 89

Table 8 Risk of bias assessment, Critical Appraisal Skills Programme (304,305) for Clinical
Decision Support in Addiction and Concurrent Disorders: A Systematic Review and Meta-
Analysis ...................................................................................................................................... 90
Table 9 Complete List of Variables Utilized ........................................................................................................ 101
Table 10 Variables included in predictive modelling for fatal overdose ......................................................... 106
Table 11 Evaluation of class balanced sample when predicting fatal overdose ................................................. 114
Table 12 10-fold Cross validation for Caret models predicting fatal overdose ................................................. 117
Table 13 10-fold Cross validation for Tidy models predicting fatal overdose .................................................. 117
Table 14 Test results from caret models for fatal overdose .............................................................................. 118
Table 15 Test results from Tidy models for fatal overdose .............................................................................. 119
Table 16 Variables included in predictive modelling for general overdose ............................................... 126
Table 17 10-fold cross validation results for caret models predicting overdose ........................................... 131
Table 18 10-fold cross validation results for tidy models predicting overdose ............................................... 132
Table 19 Test results for caret models predicting overdose ........................................................................... 132
Table 20 Test results for tidy models predicting overdose ........................................................................... 133
Table 21 Test results for Caret models predicting overdose with variables selected by Boruta. 133
Table 22 Top 10 risk factors from inference based models in Fatal Overdose Prediction .......................... 146
Table 23 Top 10 variables of importance for Overdose ............................................................................... 156
Table 24 CDR Case Definitions (338) ............................................................................................................ 254
List of Figures

Figure 1: Shift in drug markets as discussed by Krausz RM et al. (6) .......................................................... 5
Figure 2 Annualized rates of unregulated drug overdose deaths from BC Corner’s Service (16) . 6
Figure 3 Snapshot of drug overdose deaths in 2017 (16) ............................................................................. 7
Figure 4 Snapshot of drug overdose deaths in 2023 (19) ............................................................................. 8
Figure 5. Identification of Fatal and Non-Fatal Overdose Cases in the BC Overdose Cohort (43) ........................................................................................................................................... 16
Figure 6 Steps taken to create the B.C. Provincial Overdose Cohort (43) ...................................................... 18
Figure 7. Summary of the Data Components of the BC Provincial Overdose Cohort (43) .......... 19
Figure 8 ML PRISMA flow diagram (104) ........................................................................................................ 45
Figure 9 Comparative Distribution of Model Types ........................................................................................ 48
Figure 10 Comparative Distribution of Model Evaluation Metrics ................................................................. 49
Figure 11 Meta-analysis of machine learning model AUCs in cohort studies, categorized by predictive out-comes ........................................................................................................................................... 50
Figure 12 Meta-Analysis of machine learning model AUCs in cohort studies, categorized by model type .................................................................................................................................................. 51
Figure 13 CDSS PRISMA flow diagram (104) ................................................................................................. 71
Figure 14 Comparative Analysis of Publication Count and Types of Clinical Decision Support Systems by Year .................................................................................................................................. 75
Figure 15 Meta-analysis of efficacy/satisfaction & acceptance of CDSS in RCTs included in this systematic review ........................................................................................................................................ 76
Figure 16 Risk of Bias Assessment in Intention-to-Treat Analysis: ROB2 Visualization for Clinical Decision Support in Addiction and Concurrent Disorders: A Systematic Review and Meta-Analysis ........................................................................................................................................ 86
Figure 17 Data Sources Overview for the Pilot Predictive Model ................................................. 97
Figure 18 Data Sources Overview for the creation of the main dataset ........................................ 99
Figure 19 Deep learning training and validation epochs for predicting fatal overdose .............. 119
Figure 20 Area Under the Receiver Operating Characteristic (AUROC) for models predicting overdose ........................................................................................................................................ 120
Figure 21 Forest plot of AUC scores from models predicting fatal overdose ............................... 121
Figure 22 Deep learning training and validation epochs for predicting overdose ........................ 134
Figure 23 Area Under the Receiver Operating Characteristic (AUROC) for models predicting overdose ........................................................................................................................................ 135
Figure 24 Forest plot of AUC scores from models predicting overdose ........................................ 136
Figure 25 Raw AUC scores from systematic review Chapter 2 ................................................... 139
Figure 26 Raw AUC scores of models predicting opioid overdose from systematic review Chapter 2 .......................................................................................................................................... 140
Figure 27 Variable Importance in Fatal Overdose Prediction Using the Boruta Algorithm ..... 144
Figure 28 Logistic Regression Odds Ratio of variable when predicting fatal overdose .............. 145
Figure 29 XGBoost Shap Value results for Fatal Overdose Prediction .................................................... 147
Figure 30 SHAP values between total OD number and Multiple mental illnesses .................... 148
Figure 31 SHAP values between total OD number and Mood Disorders ................................. 148
Figure 32 Correlational plot for fatal overdose (overdose population) ....................................... 149
Figure 33 Variable Importance in Overdose Prediction Using the Boruta Algorithm .............. 154
Figure 34 Logistic Regression Odds Ratio of variable when predicting overdose ................... 155
Figure 35 XGBoost SHAP Value results for Overdose Prediction .................................. 157
Figure 36 Correlational plot for overdose ..................................................................... 158
Figure 37 Individual patient survival curves from RF model ........................................... 163
Figure 38 Survival curves of Cox, KM and RF models ....................................................... 163
Figure 39 Aalen's additive regression model for survival curves ....................................... 164
Figure 40 Legend for Aalen's additive regression model for survival curves ..................... 165
Figure 41 Significant Aalen's additive regression model for total overdose number .......... 165
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>AUROC</td>
<td>Area Under the Receiver Operating Characteristic</td>
</tr>
<tr>
<td>Caret</td>
<td>Classification And REgression Training R package</td>
</tr>
<tr>
<td>CART</td>
<td>Classification and Regression Trees</td>
</tr>
<tr>
<td>CDR</td>
<td>Chronic Disease Registry</td>
</tr>
<tr>
<td>CV</td>
<td>Cross Validation</td>
</tr>
<tr>
<td>DAD</td>
<td>Discharge Abstract Database</td>
</tr>
<tr>
<td>EDA</td>
<td>Exploratory Data Analysis</td>
</tr>
<tr>
<td>GBM</td>
<td>Gradient Boosting Machine</td>
</tr>
<tr>
<td>KNN</td>
<td>K-Nearest Neighbors</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>MICE</td>
<td>Multiple Imputation by Chained Equations</td>
</tr>
<tr>
<td>MSP</td>
<td>Medical Service Plan</td>
</tr>
<tr>
<td>NPV</td>
<td>Negative Predictive Value</td>
</tr>
<tr>
<td>OD</td>
<td>Overdose</td>
</tr>
<tr>
<td>OAT</td>
<td>Opioid Agonist Therapy</td>
</tr>
<tr>
<td>PharmaNet</td>
<td>A networked information service for dispensing medication in British Columbia</td>
</tr>
<tr>
<td>PPV</td>
<td>Positive Predictive Value</td>
</tr>
<tr>
<td>PWUD</td>
<td>People Who Used Drugs</td>
</tr>
<tr>
<td>R</td>
<td>A programming language and environment for statistical computing</td>
</tr>
<tr>
<td>RStudio</td>
<td>An integrated development environment for R</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
</tbody>
</table>
Glossary

**Addiction Psychiatry:** A medical subspecialty focusing on the evaluation, diagnosis, and treatment of people suffering from addiction disorders.

Risk Prediction: The use of statistical and machine learning techniques to estimate the likelihood of future events or outcomes, such as fatal overdoses.

**Artificial Intelligence (AI):** The simulation of human intelligence in machines programmed to mimic human thought processes like learning, reasoning, and self-correction.

Data Science: An interdisciplinary field focused on extracting knowledge and insights from structured and unstructured data using various techniques, processes, algorithms, and systems.

**eHealth:** The use of information and communication technologies for health.

**Fatal Machine Learning (ML):** A subset of artificial intelligence involving the development of algorithms that enable computers to learn from and make predictions or decisions based on data.

**Fatal Overdose:** An acute adverse event resulting in death due to the consumption of excessive amounts of a substance, often related to drugs or medication.

General Overdose: An overdose event where the individual consumes an excessive amount of a substance but survives. It may still result in significant harm or require medical intervention.

**Opioid Agonist Therapy (OAT):** A treatment for opioid addiction that uses medications, such as methadone or buprenorphine, which activate the same opioid receptors in the brain as the addictive substance but are safer and less likely to produce harmful behaviors.

**PWUD (People Who Use Drugs):** A non-stigmatizing term in public health and addiction psychiatry referring to individuals who consume drugs, emphasizing their personhood over their drug use.
**Sample:** In the context of research, a subset of data or population used for analysis. In machine learning, this term often refers to individual data points or examples from which the model learns.

**Supervised Learning:** A type of machine learning where the model is trained on labeled data.

Statistics: The study of the collection, analysis, interpretation, presentation, and organization of data.
Acknowledgements

I would like to express my profound gratitude to all the individuals who have played an instrumental role in the completion of this project.

Firstly, I extend my deepest appreciation to the frontline healthcare workers, physicians, and various healthcare team members in BC who have been courageously battling the drug toxicity crisis over the past decade. My gratitude extends to the members of the British Columbia Centre for Disease Control (BCCDC), including Dr. Amanda Slaunwhite, Chloe Xavier, Bin Zhao, and Kevin Hu. Their dedication and efforts in overseeing the use of critical data have been indispensable. I am also grateful to Craig Hutton who works for the Province of British Columbia's Behavioural Insights Group, for his insights and contributions to this data. A special thanks to Tim Choi, the lead for data partnerships from Population Data BC, for his invaluable support.

I am deeply grateful to my supervisor, Dr. Krausz, for his pivotal role in my professional development, offering me countless opportunities to enhance my skills in research, academia, and beyond. I would also like to extend a special thank you to Dr. Kazemi, who has been instrumental in shaping the vision of this project from its inception, guiding me with invaluable insights and direction. My gratitude also goes to Dr. Kerry Jang, whose mentorship has significantly influenced my growth, writing skills, and attention to detail. Furthermore, I am thankful for the involvement and support of Dr. Raymond Ng and Dr. Christian Schutz in my project. Their expertise and contributions have been greatly appreciated.
Working with the Addiction and Concurrent Disorders Group at UBC, and UBC colleagues has been an enriching experience, and I am thankful for the collaborative efforts of Janet Suen, Jane Kim, Vanessa Kitchin, Mostafa Mamdou, Farhud Shams, Jean Westenberg, James Wong, Mohammad Ali Nikoo, Marisha Boyd, Tam To, and Fiona Choi. Their support and teamwork have been crucial to the success of this research.

I extend my heartfelt thanks to our international collaborators, including Dr. Max Meyer from the Universitäre Psychiatrische Kliniken Basel and Dr. Jim Schmeckenbecher from the Medical University of Vienna, for their support in my research. Their expertise and insights have been fundamental to its success. Additionally, I am profoundly grateful to Dr. Ian Hickie, Dr. Frank Iorfino, and Dr. Mathew Varidel at the University of Sydney for providing me with the exceptional opportunity to complete an international internship. Their guidance and support have been instrumental in broadening my academic and professional horizons.

I would like to acknowledge the UBC programs in Computer Science and Master of Data Science, which provided me with teaching opportunities. I am thankful to professors Dr. Vincent Vincenzo, Professor Tiffany Timbers, and Professor Joel Ostblom for their guidance and support.

Lastly, my heartfelt thanks go to my family and friends for their unwavering love and support. This includes Jason Edgar, Farhud Shams, my parents, Joe and Paulina Tai, my dear sister Alicia Tai, and my grandmother Chan Kwei To. Your encouragement and belief in me have been my constant source of strength.
Dedication

To beautiful Vancouver, my hometown that has nurtured my dreams and ambitions. To the resilient community that thrives amidst its urban beauty, yet silently battles the shadows of hardship. This work is dedicated to the souls touched by the drug toxicity crisis—a crisis that indiscriminately weaves through the fabric of our society, leaving a tapestry of loss and hope.

In the memory of those we've lost, and in the spirit of those who continue to fight, this dedication stands as a beacon of hope. For a community united in overcoming adversity, and for a future where every individual can thrive, free from the chains of addiction.
Chapter 1: Introduction

1.1 Institutional change and the rise of homelessness in Vancouver

In the vibrant heart of Vancouver, British Columbia (BC), Canada, a complex and troubling story of homelessness and substance abuse has been unfolding for decades. This narrative, deeply embedded in the cityscape of one of Canada's most scenic urban centers, is rooted in a series of systemic changes and policy decisions that have profoundly affected vulnerable groups. A critical turning point in this ongoing challenge of homelessness and addiction was the closure of Riverview Hospital, a major mental health institution in Coquitlam, in the 1990s (1).

This closure was part of a broader shift in attitudes towards mental health care. Across the globe, institutions like Riverview faced criticisms for issues such as underfunding, overcrowding, and the alienation and isolation of patients from society (2). In response to these criticisms, BC experienced a significant move towards deinstitutionalization of mental health care between 1980 and 1990. However, the operationalization and practical outcomes of this shift appeared to be inadequately considered and poorly executed.

Riverview Hospital, once a cornerstone of mental health care, represented a traditional approach to psychiatric treatment. Its closure, driven by shifting attitudes towards mental health care and financial limitations, led to unforeseen repercussions. This decision displaced many patients, who still required ongoing psychiatric support, at a time when the urban environment was undergoing significant changes. In a situation reminiscent of developments in places like Verona, Italy, and Sheffield, United Kingdom, Vancouver struggled with inadequate coordination across different mental health service levels, a lack of legislative commitment and
political backing, funding issues, and insufficient community-based mental health services (3). These factors failed to adequately offset the reduction in centralized institutional capacities. As a result, regional general hospitals and community-based services faced overwhelming demands. Moreover, the lack of funding and legislative commitment hindered effective transition of care for patients from Riverview to community settings, despite the provincial government's unfulfilled promises of support.

In Vancouver, the lack of effective coordination among different levels of mental health services, the absence of strong legislative commitment and political backing, as well as funding challenges, meant that the community-based mental health services were inadequately prepared for the reduction in centralized institutional capacities. Consequently, regional general hospitals and community services found themselves under significant strain. The scarcity of funding and legislative support also erected considerable barriers in transitioning patients from Riverview to community care. Promises made by the provincial government for sufficient housing for these individuals largely went unfulfilled, exacerbating the challenges faced in the transition period.

In the wake of limited housing options, many patients from Riverview Hospital sought refuge in the affordable hotels of Downtown Eastside, implemented as a part of the (4). This area, already struggling with socio-economic challenges, became a hub for vulnerable individuals, including those displaced from the closed mental health facility. These groups, vulnerable and often battling existing mental health comorbidities, became easy targets for drug dealers, leading to an escalation in drug addiction. This amalgamation of factors – the influx of a vulnerable burdened with physical and mental comorbidities, the transformation of hotels into low-income housing with subpar conditions, and the pre-existing poverty – laid the groundwork for a burgeoning public health crisis. The introduction of opioids into this mix further intensified
the situation, making Downtown Eastside a hotbed for the drug toxicity crisis that has swept across North America. In this environment, individuals facing both mental health challenges and homelessness found themselves particularly susceptible to addiction, exacerbating the already dire circumstances and highlighting systemic failures in addressing the complex nexus of mental health, addiction, and homelessness (5).

Compounding this crisis were various contributing factors. Economic disparities, changes in drug availability and composition, insufficient mental health support post-deinstitutionalization, ineffective policing and drug policies, lack of adequate addiction services and harm reduction measures, and societal stigma around addiction all played significant roles. Additionally, social isolation, community disintegration, limitations within the healthcare system, and Vancouver's position in the international drug trade further aggravated these challenges.

The healthcare system struggled to cope with this escalating crisis. Constraints in providing adequate pain management and addiction treatment, combined with the broader complexities of the healthcare landscape, contributed significantly to the misuse of opioids. This scenario resulted in a distressing spike in opioid-related fatalities, underscoring the urgency for a comprehensive and coordinated approach to tackle this multifaceted emergency, an emergency that has continued to grow more acute with time.

1.2 The emergence of opioids

A recent publication outlined the emergence of the drug toxicity crisis in North America as occurring in three distinct phases (6). Initially, the crisis was driven by a marked rise in the prescription of opioids such as OxyContin for pain management and post-surgical care, leading
to some of the highest per capita opioid usage rates in the United States and Canada (7–9). The
second phase emerged around 2010, following a reduction in these prescriptions, which led to an
increase in the use of street heroin among those addicted to prescription opioids (10). The third
phase commenced in 2013, with the introduction of synthetic opioids into the drug market (6,10).

British Columbia (BC) is currently grappling with an drug toxicity crisis of
unprecedented scale, a crisis that has evolved significantly over the past decade. Traditionally,
discussions around opioid-related issues have often centered on prescription opioids. However,
the current landscape in BC tells a different story, one where the problem extends far beyond the
realm of prescription medications. Initially, the rise in opioid use and subsequent addiction rates
were closely linked to prescription opioids. These medications, intended for pain management,
were often misused, leading to dependency and, in some cases, overdose(11,12). In response,
healthcare providers and policymakers implemented stricter prescription guidelines and
monitoring systems to curb this trend (13). These measures, while effective in reducing
prescription opioid misuse, inadvertently pushed some individuals towards illicit opioids.
Contrary to other regions in North America, during the onset of BC’s drug toxicity crisis, a
significant proportion of individuals who overdosed did not possess a prescription for opioid
pain treatment (14).
Figure 1: Shift in drug markets as discussed by Krausz RM et al. (6).

Highlighted quite well by Krausz et al., the drug toxicity crisis in BC has undergone significant evolution, beginning with the limited presence of fentanyl, primarily in diverted medical patches, before 2013 (6). This initial stage was markedly different from Europe's continued limited use of fentanyl. A notable shift occurred as fentanyl, mostly imported from China, began contaminating the street drug market, leading to a rise in overdoses due to users being unaware of its presence in heroin (15). The period from 2013 to 2018 saw changes in substance use patterns, with a notable increase in inhalation methods in BC (6). Ultimately, fentanyl started to take over the market. For instance, the BC Coroner’s Service reports that a majority of the drug toxicity deaths in recent years involved fentanyl, as opposed to prescription opioids with illegal fentanyl detected in 60% of deaths as unregulated drug overdoses at the start of the crisis in 2016 (16).
Post-2018, fentanyl became omnipresent in street drugs, with policies targeting international trafficking. China's 2019 crackdown on fentanyl production led to shifts in the drug trade, including the importation of precursors for local production in North America and Mexico (17). The trade through Mexico and the increase in domestic production in BC, especially post-2019, resulted in a more potent and dangerous drug supply, significantly impacting overdose rates (18).

1.3 Current drug toxicity crisis

In April 2016, British Columbia (BC) declared a public health emergency as seen in figure 2 and figure 3, in response to increased mortality from opioid overdoses (16). Despite the concerted efforts of the BC health authorities, the past four years have witnessed a significant and concerning increase in the number of deaths due to overdoses. The statistics reflect a concerning trend: 1,734 deaths in 2020, rising to 2,269 in 2021 (19). The upward trajectory continued with 2,377 fatalities recorded in 2022 (19). By October 2023, the figures have already reached 2,039, indicating a persistent and troubling issue as seen in figure 4 (19).

Figure 2 Annualized rates of unregulated drug overdose deaths from BC Corner’s Service (16)
Continuing the trend observed since 2016, fatalities involving synthetic opioids, particularly fentanyl, have surged. Fentanyl use has only drastically increased to 86% in 2021, and 82% in 2022 (20). It is essential to challenge the misconception that the drug toxicity crisis in BC is primarily due to the misuse of prescription opioids. Such a view underestimates the complexity of the issue, especially in light of the growing prevalence of illicit synthetic opioids like fentanyl. The crisis demands a broader perspective that encompasses not only the types of substances used but also the socio-economic elements contributing to addiction, the mental and physical health comorbidities, and the critical need for adequate housing (21). These comprehensive factors are integral to understanding and addressing the multifaceted nature of drug use and addiction in BC.
The homeless population in BC, already in a precarious situation, contends with a multitude of comorbidities (22). These include addiction and mental illness, which are further compounded by physical health challenges. These health issues often disproportionately impact specific ethnic groups, reflecting broader systemic inequalities (23). Additionally, this demographic is frequently characterized by low socioeconomic status, which can exacerbate their health challenges and limit access to essential resources.

A particularly pressing concern in this context is the limited access to adequate treatment and professional healthcare. This lack of access is not merely a matter of physical availability of services; it is also influenced by barriers such as stigma, lack of awareness about available resources, and the absence of tailored healthcare approaches that address the unique needs of the homeless. The intersection of these factors creates an environment where preventive care is
scarce, and early intervention is rare, leading to a higher incidence of medical crises, including overdoses.

Moreover, within the homeless population, there is a significant variation in the level of healthcare needs. Some individuals may require intensive, long-term support due to chronic conditions, while others might benefit from more immediate, crisis-oriented interventions. The challenge lies in the healthcare system's ability to provide a flexible, responsive range of services that can adapt to these diverse needs.

The exacerbation of overdose numbers among the homeless is a stark indicator of these systemic failures. It underscores the need for a holistic approach to healthcare that is not only accessible but also culturally sensitive, stigma-free, and attuned to the complex realities of those living without stable housing. Such an approach should ideally integrate physical health services with mental health support and addiction treatment, thereby offering a comprehensive care model that can effectively address the multifaceted nature of health challenges faced by the homeless.

1.4 Current healthcare approach

In Canada, the failure of the system of care is measured by different indicators (24). These indicators include low Opioid Agonist Treatment (OAT) coverage and low needle-syringe distribution rates, compared to the rest of the world (24). In recent years BC has tried to adopt different global strategies to tackle the drug toxicity crisis.

For the last 3 years, the BC government is focusing on the 'safe supply' strategy, which involves prescribing potent substances outside of a therapeutic context (25). This approach is primarily aimed at addressing the increased toxicity in drug supply due to the contamination with
synthetic substances like fentanyl. While it is reasonable to associate the surge in overdose deaths with the prevalence of fentanyl, the BC government's response appears to be somewhat impulsive, reminiscent of the hasty deinstitutionalization of Riverview. This suggests a pattern of actions taken without fully comprehending their broader implications. With recent evidence, it is clear that BC’s safer supply policy is not only ineffective, but associated with more opioid related poisoning hospitalizations (26).

Though well-intentioned, the focus on harm reduction represents only a single aspect of the comprehensive four-pillar approach modeled by Switzerland. The Swiss model, which has shown remarkable success in mitigating substance abuse issues, includes prevention, treatment, harm reduction, and enforcement (27). These pillars are designed to work synergistically, offering a balanced and multifaceted approach to substance abuse.

By concentrating primarily on harm reduction through the safe supply initiative, BC seems to overlook the critical importance of the other three pillars. The absence of a strong emphasis on prevention, treatment, and enforcement means that the approach is less holistic than the Swiss model. This partial adoption could limit the effectiveness of BC's strategy in addressing the complex and multifaceted nature of substance abuse and addiction. The success of Switzerland's strategy lies in its integrated approach, something that BC needs to consider for a more impactful and comprehensive response to its ongoing drug crisis. Among the three critical issues, the most alarming is the substantial gaps in the Canadian healthcare system. This includes a notable lack of treatment options (28), limited access to mental healthcare and psychosocial interventions, and often lengthy wait times. These factors have been identified as significant barriers to retaining patients in Opioid Agonist Therapy (OAT) (29). Recent data indicates an average wait time of 24.4 weeks from general practitioner referral to treatment initiation in 2019.
Furthermore, there is currently no comprehensive review of the 'safe supply' approach. A rapid review conducted in 2022, which covered 19 studies, found no clear benefits in providing pharmaceutical opioids, heroin, crystal methamphetamine, cocaine, or similar substances to individuals with dependencies on these drugs (31). However, when examining the available treatments within BC’s healthcare system, the limitations become even more apparent, especially when compared to the more diverse OAT employed in Switzerland.

In BC, the primary Opioid Agonist Treatments (OAT) include methadone, buprenorphine/naloxone (Suboxone), and slow-release oral morphine (Kadian) (32). This contrasts with Switzerland's broader range of OAT options, such as codeine, dihydrocodeine, hydromorphone, and diacetylmorphine, along with experimental treatments like intranasal diacetylmorphine (33,34). Despite Health Canada's approval of injectable hydromorphone and diacetylmorphine in 2019, the expansion of substitution clinics that offer these treatments beyond the crosstown clinic in Vancouver has been limited. Research in BC indicates that less than 60% of OAT episodes completed the initial induction among 45,608 individuals between 2008 and 2017 (32). Shifting focus from the specific Opioid Agonist Therapy (OAT) options available in BC to the broader spectrum of treatment methodologies, a significant contrast emerges in the management of drug crises between BC and Switzerland, particularly regarding the provision of ongoing and supportive care following an overdose incident.

A striking difference in the treatment trajectories between BC and Switzerland lies in their holistic approach. Unlike BC, Switzerland's approach ensures vulnerable individuals are not returned to the streets immediately after being revived from an overdose, demonstrating a more comprehensive and supportive treatment model. When comparing the treatment trajectories between BC and Switzerland, the incorporation of therapy, psychosocial support, peer-based
programs, and housing stands out in Switzerland's approach. Swiss OAT is not just about medication; it encompasses a holistic model where psychosocial support plays a significant role. This includes counseling, group therapy, and peer support programs that aid in social reintegration and mental health support. Additionally, housing initiatives are often integrated, recognizing the crucial role of stable living conditions in recovery and overall well-being. Such a comprehensive approach contrasts with BC's more medication-focused strategy, highlighting the need for broader support services alongside medical treatment for opioid dependency.

In conclusion, the challenges in addressing the drug toxicity crisis in Vancouver are multifaceted, stemming from prolonged wait times and difficult access to care, limited treatment options, and a lack of comprehensive care trajectories that include prevention, treatment, monitoring, and management. These issues are compounded by healthcare shortages and may be linked to funding constraints. In response to these challenges, the BC Centre for Disease Control (BCCDC) established the BC Provincial Overdose Cohort, a database that has been instrumental in enhancing our understanding of the risk factors contributing to overdoses. This cohort has opened new avenues for research and the development of targeted solutions to support the vulnerable population.

Transitioning from the challenges to the potential solutions, the BC Provincial Overdose Cohort serves as a pivotal tool in the fight against the drug toxicity crisis. By leveraging the data gathered in this cohort, researchers and policymakers can identify key areas for intervention and develop strategies that are not only reactive but also proactive in nature. The next section will delve into how the insights derived from the BC Provincial Overdose Cohort are shaping innovative approaches and policies aimed at mitigating the drug toxicity crisis and providing more effective support for those at risk.
1.4.1 Overdose population

In BC, the drug toxicity crisis has significantly impacted a diverse population. From 2013 to 2023, there's been an alarming trend in unregulated drug deaths, especially among 19–59 year olds (19). October 2023 alone saw 189 deaths, indicating a fluctuating crisis. Males, particularly in the 30-59 age bracket, represent most of these deaths (19). Geographically, Vancouver, Surrey, and Greater Victoria, along with Vancouver Coastal and Fraser Health Authorities, experience the highest fatalities (19). Most deaths occurred indoors, with smoking as the most common method of consumption (19). Fentanyl's role in these deaths has grown significantly, alongside other substances like cocaine, methamphetamines, and benzodiazepines (19). This trend underscores the escalating complexity and severity of the drug crisis in BC, spanning various demographics and communities.

One group that has been disproportionately affected is the First Nations people due to the ongoing legacy of colonization. First Nations peoples in Canada face various forms of racism, including interpersonal, systemic, and structural, which act as barriers to accessing appropriate care. This discrimination, coupled with the trauma from oppressive colonial policies like residential schools, leads to distrust in the healthcare system (35,36). The residential school system, marked by abuse and forced relocation, contributed significantly to historical trauma, affecting generations with feelings of shame, loss, and self-hatred (37). This intergenerational trauma is linked to higher rates of substance use, mental health issues, and limited access to treatment among First Nations peoples (38).

Despite constituting just 3.4% of BC's population, First Nations people disproportionately experienced 14% of all overdose events in the province in 2016 (39). Alarmingly, they were five times more likely to suffer an overdose compared to non-First
Nations individuals, highlighting a significant disparity in the impact of the overdose crisis on this community (39). Ten percent of all overdose deaths in BC were First Nations and they were three times more likely to die from an overdose (39).

1.4.1.1 Overdose prevention

In 2023, while BC has taken significant steps in overdose prevention through the implementation of safe injection sites, drug checking with fentanyl test strips, and First Nations treatment centers, there are still critical areas that require further development. The absence of comprehensive peer support groups is one such gap (40). Peer groups can play a vital role in providing support, sharing lived experiences, and fostering a community of care and understanding, which is especially crucial for long-term recovery and relapse prevention (41).

Additionally, there is a need for more robust education and prevention programs. These programs are essential not only for at-risk individuals but also for the broader community to understand the complexities of substance use disorders and the challenges faced by those struggling with addiction. Effective education programs can help in destigmatizing drug use, spreading awareness about harm reduction practices, and informing people about the resources available for support and treatment.

Overall, while BC's measures in addressing the overdose crisis are commendable, the addition of peer-led support groups and enhanced education and prevention initiatives could provide a more comprehensive and effective approach to tackling the challenges of drug use and overdoses. A descriptive analysis conducted on data from the BC Provincial Overdose Cohort suggests there may be opportunities to identify individuals at risk of overdose before it happens.
Utilizing this data effectively could significantly enhance these efforts by enabling earlier identification and intervention for individuals at risk.

1.5 British Columbia Provincial Overdose Cohort

The British Columbia Provincial Overdose Cohort is a population-level dataset collected by the British Columbia Centre for Disease Control (BCCDC) of all identified persons who experienced a drug-related overdose in BC, Canada, between January 1st 2015 to December 31st, 2019 (16).

Individuals who suffered an overdose were pinpointed using multiple data sources, including BC Emergency Health Services, the Drug and Poison Information Centre, the BC Coroner’s Service, case reports from Emergency Departments within the Interior Health Authority, Vancouver Island Health Authority, and Northern Health Authority (each providing three distinct files), as well as the National Ambulatory Care Reporting System, the Discharge Abstract Database, the Medical Services Plan, and the Vital Statistics Deaths registry (43,44). In response to a public health emergency declared because of a sharp rise in opioid-related general and fatal overdoses, the BC Provincial Health Officer directed all Emergency Department (ED) administrators and physicians to report overdose cases to the Medical Health Offices (43). The Interior Health Authority (IHA), Vancouver Island Health Authority (VIHA), and Northern Health Authority (NHA) lacked dedicated systems to track opioid-related ED visits (43). Medical staff in these regions were requested to fill out a manual case report form for each overdose-related ED visit, suspected or confirmed to involve opioids or other drugs, and submit these to their local Medical Health Office (43). These reports included limited contextual details.
about the overdose incidents. An assessment of the VIHA's case reports indicated a significant underreporting of ED visits for opioid overdoses, with a discrepancy rate of 55% (43).

Apart from identifying overdose, this Cohort integrates several distinct datasets, offering insights into various potential factors contributing to overdoses. It is recognized as the most extensive repository of data concerning overdose incidents and associated risk factors within BC. Figure 5 illustrates the formation of this Cohort, which relies on data passively collected from healthcare services in BC. It includes administrative data, routinely gathered by public and government entities. However, this dataset has limitations in addressing certain research queries. For instance, it cannot account for overdose events where individuals did not seek healthcare services, hence not capturing every overdose occurrence in the province.

Figure 5. Identification of Fatal and Non-Fatal Overdose Cases in the BC Overdose Cohort (43)

1.5.1 An overview of the data

The BC Overdose Cohort, managed by the British Columbia Centre for Disease Control (BCCDC), integrates data from a broad range of sources, including the Ministry of Health, the
Provincial Health Officer’s Order, and various Health Authorities under the BCCDC (43). The BCCDC Provincial Overdose Cohort Team, in collaboration with the Provincial Health Services Authority (PHSA) Decision Support Services Teams, collected additional datasets not initially provided by the Ministry of Health (MOH) (43). These included data from the BC Emergency Health Services (BCEHS), Emergency Departments (ED), BCCDC Vital Statistics, BC Coroner’s Service (BCCS), and the Drug and Poison Information Centre (DPIC) (43). These were combined with personal identifiers and sent to the MOH for data linkage. The MOH then merged these with its datasets, comprising the Client Registry (CLR), Chronic Disease Registry (CDR), Medical Service Plan (MSP), Discharge Abstract Database (DAD), PharmaNet (PNet), Vital Statistics Deaths (VSD), and National Ambulatory Care Reporting System (NACRS), to further identify potential drug-related overdose cases (43). The PHSA also contributed BC Corrections data and the First Nations Client File (FNCF) to the Cohort. After the integration of all datasets, the MOH prepared a de-identified version of the Cohort for preliminary data cleaning by the PHSA Decision Support Services Team (43). This de-identified Cohort was then forwarded to the BCCDC Provincial Overdose Cohort Team for additional cleaning, processing, and organizing into analysis-ready files (43). Figures 6 and 7 provide visual representations of this process and the categorization of the datasets by source.

The data within the BC Overdose Cohort is systematically organized into four primary categories, as illustrated in Figure 7: Population data from the PHSA, Healthcare data, Emergency services data, and Death records. This comprehensive categorization ensures a holistic approach to data analysis and interpretation within the scope of the Cohort's objectives.

This Cohort also contains extensive data, encompassing overdose events, prescription medications, social assistance program engagement, mental health service utilization, provincial
incarceration history, and healthcare usage, including hospitalizations and emergency department visits (43). The health data in the Cohort extends back to 2010 and receives annual updates for continuous expansion and relevance (43). Such a longitudinal design enables an in-depth examination of events preceding, during, and following an overdose (43). Health Authorities involved in the overdose response are permitted to request data access, making this Cohort a crucial tool for provincial-level research, and offering critical insights into a broad range of datasets (43).

Figure 6 Steps taken to create the B.C. Provincial Overdose Cohort (43)
1.5.2 Defining Cases and Controls in the Overdose Cohort

The Cohort includes overdose cases and controls (43). Cases refer to individuals in the Cohort who experienced a drug-related overdose since January 1, 2015, a period marking a significant increase in illicit drug overdoses and deaths in BC (43). An overdose event is identified by one or more of the following: naloxone administration by paramedics, a call regarding an opioid-related event to the Drug and Poison Information Center, a physician's diagnosis of an opioid overdose in an emergency department, a coroner's determination of illicit drug toxicity death, a hospital or emergency department visit with an overdose diagnosis code, or a Vital Statistics overdose-related death (43). To avoid duplicate counting, overdose events
recorded in multiple datasets, like ambulance, emergency, and hospital records, are consolidated (43).

Controls, on the other hand, serve as a comparison group for the Cohort. They are selected from a 20% random sample of the BC population, ensuring access to their health and prescribing histories (43). A matched control group is also utilized, typically aligned with the Cohort in terms of birth year, sex, and Local Health Area of residence for more precise comparisons (43).

1.5.3 The integration of the BCCDC dataset for a E-health platform, a clinical decision support system

The BC Provincial Overdose Cohort, with its extensive data on drug-related overdoses, presents a unique opportunity for integration into an E-health online mental health platform. This integration can be transformative, harnessing the depth and breadth of the Cohort's data to enhance the effectiveness of digital mental health interventions. By leveraging the detailed information on overdose events, prescription medications, and healthcare utilization, the platform can tailor its services to the specific needs and risk profiles of individuals. This data-driven approach enables the identification of high-risk individuals and the provision of targeted, timely mental health support.

Building on this foundation, a Clinical Decision Support System (CDSS) integrated into the E-health platform can act as a powerful tool to distill and apply the Cohort's data effectively. A CDSS analyzes complex healthcare data to provide evidence-based recommendations tailored to the individual's health profile. In the context of the BC Provincial Overdose Cohort, it means that the CDSS can help clinicians identify subtle patterns in patient behavior, prescription usage,
and past overdose incidents that might go unnoticed by the human eye. Such insights are crucial for early intervention and can inform clinicians about the optimal timing for intervention, adjust treatment plans, and predict patient trajectories. Moreover, the CDSS can alert healthcare providers to emerging trends, such as an uptick in overdose rates, enabling a swift and coordinated response. The integration of a CDSS transforms static data into dynamic, actionable insights, elevating the role of E-health platforms in preventing overdoses and managing addiction, thus marking a significant step towards an intelligent, data-informed approach to public health in BC.

### 1.6 Risk Assessment and Management Platform

#### 1.6.1 E-Health Model

Web-based interventions are increasingly recognized for their accessibility, efficacy, and cost-effectiveness in promoting health behavior change, surpassing many traditional treatments in these aspects (45–47). Notably, internet interventions have proven effective in reducing opioid, cocaine, and amphetamine use, as evidenced by a meta-analysis by Boumparis et al. and further supported by Marsch et al.'s 2014 study (48,49). These studies highlight the enhanced effectiveness of web-based behavioral interventions when combined with medication-assisted treatment.

A significant advantage of web-based interventions is their suitability for delivering a continuum of care, thereby improving coordination and service delivery across various needs. For example, the eCLIPSE platform in Australia successfully integrates online and face-to-face support for individuals with concurrent mental health and substance use disorders (50). Such
web-based tools, incorporating diverse psychosocial therapeutic approaches, have shown efficacy in reducing opioid use and bridging systemic gaps.

Different online tools, including web-based risk assessment scales paired with tailored health advice, have demonstrated effectiveness in initiating health behavior change and mitigating various health risks, such as cardiovascular disease (51,52), diabetes (53,54), pressure ulcers (55), and HIV (56,57). The ORION E-health tool, specifically targeting overdose risk, exemplifies this by providing a reliable evaluation of overdose risk and promoting engagement and willingness among individuals to modify their risk factors (58). Another tool that is interesting is the predict model that helps with depression medication (59,60). There also exists models in literature that can predict risk associated with substance use and addiction related outcomes (61,62). Chapter 3 delves into significant clinical decision support systems such as the one RAMP uses, which incorporate these models for clinical application in addictions and concurrent disorders.

The inclusion of goal setting and feedback in E-health interventions have also been shown to enhance behavioral change. This is evident in the reduction of substance use days and improved treatment initiation. Moreover, peer support plays a crucial role in substance use treatment engagement and related behaviors, such as craving and self-efficacy. Peers can bridge the gap between public health workers and People Who Use Drugs (PWUD), especially crucial given the latter's often limited access to care (63).

While concerns about PWUD's access to technology and digital literacy exist, studies indicate relatively high technology access among this group, with a significant interest in online health interventions. Preliminary survey results from our research team in Vancouver indicate
high levels of internet access, smartphone ownership, and interest in using technology for health management among patients with substance use and concurrent disorders.

The acceptability of online interventions, as illustrated by the success of models like Innowell’s Project “Synergy” from the University of Sydney, underscores the vast potential of web-based solutions for PWUD (64). Project Synergy was a government-funded series of research trials aiming to assess the effectiveness of the Synergy Online System, a digital platform for delivering mental health care (64). This comprehensive project, spanning various demographic groups including children, young people, adults, and the veteran community, sought to measure the impact of this technology in Australian mental health services. This project exemplifies how digital solutions can revolutionize healthcare, especially for vulnerable groups such as PWUD. By integrating such innovative E-health models, similar to Innowell’s approach, there is an opportunity to significantly enhance technological access and the overall effectiveness of interventions in the context of addiction and mental health. This aligns with the broader shift towards digital solutions in healthcare, demonstrating the efficacy and scalability of such platforms in addressing complex health challenges.

1.6.1.1  E-health in Canada

In Canada, the evolution of E-health, exemplified by Alberta's virtual Opioid Agonist Treatment (OAT) clinics, represents a significant step forward in healthcare delivery (65). These virtual clinics are a testament to how technology can bridge gaps in healthcare accessibility, particularly for remote or underserved populations. They provide an innovative model for offering continuous and comprehensive care, leveraging digital platforms to extend the reach of essential health services.
The potential for applying similar virtual care systems in BC, especially in cities like Vancouver, is immense, particularly in tackling the drug toxicity crisis. By facilitating easier and more convenient access to OAT and other addiction treatments, these virtual platforms could significantly lower barriers to care (66). This approach could enhance the effectiveness of interventions, offering a more inclusive and responsive healthcare system for those struggling with opioid dependency. This integration of virtual and traditional healthcare models could be a game-changer in managing public health challenges.

Expanding E-health services in BC to include a variety of treatments can revolutionize the approach to addiction care. Beyond OAT, there's potential for online peer-led treatments, which can provide vital support and shared experiences crucial for recovery and relapse prevention (67). Additionally, virtual platforms can facilitate continuous monitoring and follow-up care, ensuring a comprehensive treatment trajectory. This approach can foster a more supportive, accessible, and efficient system, addressing the full spectrum of needs for individuals in their recovery journey.

1.6.2  Risk Assessment and Management Platform

In 2017, the Addiction and Concurrent Disorders Group (ACD Group) proposed the innovative idea of leveraging technology to develop a web-based platform. This platform aimed to empower individuals to manage personal risk factors for an overdose and to provide or facilitate access to appropriate risk management resources. Following this proposal, in 2018, the ACD Group was awarded a grant by Health Canada's Substance Use and Addictions Program.

The resulting project, RAMP, is a multi-year initiative developed to assist People Who Use Drugs (PWUD) e.g. by identifying and managing their risk of overdose through its web-
based platform. RAMP's mission is to empower individuals in managing their personal risk factors for overdose and to facilitate access to appropriate risk management resources. Created in direct response to the escalating overdose crisis, RAMP's objective is to aid both individuals and their networks in identifying key overdose risk factors and developing strategies for their management. The platform is designed to be enriched by diverse lived experiences and aims to engage users, motivating them to re-evaluate high-risk aspects of their lifestyles and substance use, offering alternative options and a deeper understanding of existing support structures.

RAMP's core functionality revolves around online risk management tools designed to enhance care access and promote the sharing of experiences and advice drawn from personal encounters. As the platform's user base expands, RAMP provides valuable insights into the prevalence and dynamics of various overdose risk factors, potentially becoming a pivotal element in unifying the healthcare system's overdose response and monitoring outcomes.

RAMP actively contributes to an effective response to Canada's overdose crisis by offering a range of comprehensive services. It provides detailed individual overdose risk assessments with standardized feedback for PWUDs, encouraging them to share these findings within their networks. This approach not only empowers the individuals directly affected but also extends the understanding and management of overdose risks to their wider community. Additionally, RAMP offers vital resources for peers and family members, enabling them to understand overdose risks and the necessary steps to mitigate them. This is complemented by training to equip them with essential knowledge and skills. Such resources are crucial for building a supportive environment around those at risk. Moreover, RAMP serves as a platform for sharing lived experiences through user-generated media. This feature is crucial in shaping the approach to creating solutions, as it provides real-world perspectives and insights. Finally,
RAMP's online resources aim to enhance behavioral skills through methods like goal setting and various behavioral management techniques, including graded tasks and contingent rewards. These resources are instrumental in promoting behavioral change and risk reduction, underscoring RAMP's multifaceted role in addressing the overdose crisis in Canada.

The platform also plays a pivotal role in developing a dynamic understanding of overdose risk factors, their significance, and how they evolve within the population. This insight is key to fostering the development of predictive models, providing a proactive strategy in the fight against the overdose crisis.

The objectives of the RAMP project are multifaceted and currently aim to engage at-risk substance user populations, enhance knowledge about overdose risks, and reduce these risks both at individual and population levels. Additionally, RAMP seeks to connect individuals to both online and physical services and establish a dynamic knowledge exchange system that caters to a wide range of stakeholders including users, families, peers, and professionals.

Presently, RAMP offers a variety of functionalities accessible through desktop and mobile platforms. These include an initial engagement portal, psychoeducation, risk assessment, goal-setting, risk management resources, and a back-end data/research platform. These components are designed to be interconnected yet distinct, with some of them requiring the creation of a user account for access to more personalized features such as goal-attainment monitoring. In summary, RAMP is envisioned as a comprehensive online platform that aims to lower the threshold for accessing various components. Future expansions are anticipated to include additional online functionalities and professional support, predicting other comorbidities such as trauma and suicide risk, and enhancing the platform's capability to effectively address the needs of its users.
1.6.3 High Risk Assessment

The development of the High Risk Assessment tool was a significant achievement by James Wong, a Master's student whose research culminated in a thesis that successfully passed its defense (68). This pilot study focused on the High Risk Assessment (HRA) primarily aimed to investigate the risk and protective factors for general opioid overdose among individuals using street fentanyl, particularly during the COVID-19 pandemic. The study was designed with three primary objectives in mind. First, it aimed to assess the prevalence of overdose incidents. Secondly, it sought to explore the relationships between known factors and the occurrence of overdoses. Finally, the study aimed to examine the links between unique factors related to fentanyl and COVID-19, and their impact on the number of overdose cases (68).

The methodology involved creating survey questions on overdose risk and protective factors. This process was iterative, incorporating a thorough review of existing literature on opioid overdose and incorporating the expertise of six clinician-scientists/clinicians who specialize in the care of People Who Use Drugs (PWUD) in BC. This robust approach informed the development of a comprehensive risk tool integrated into RAMP, designed to help individuals identify their key risk factors contributing to general overdose.

In a pursuit to enhance this tool, this project adopted a different approach, focusing on predicting the risk of both fatal and general opioid overdose. This new direction employed data-driven machine learning techniques, representing a shift from traditional rule-based methodologies to a more sophisticated, algorithmic approach. The goal was to refine the prediction of overdose risks, leveraging the vast potential of machine learning to analyze complex, multifaceted data sets.
1.6.3.1 Overdose risks

The research spearheaded by Wong, utilizing data from his cross-sectional pilot study, sheds new light on the landscape of opioid overdose risks in the current fentanyl era. The study indicates that certain risk factors from the period before the widespread emergence of fentanyl, such as being male, having a history of previous overdoses, and experiencing suicidal thoughts, remain crucial indicators of an increased likelihood of experiencing a general opioid overdose.

Intriguingly, it was observed that other factors previously thought to be significant, like the method of drug consumption and engagement in opioid agonist therapy (OAT), did not show a strong connection with the likelihood of a fentanyl-related overdose (68). Additionally, the study explored the effectiveness of the ‘safe supply’ approach, which was introduced to counteract the risks posed by the increasingly toxic drug supply during the COVID-19 pandemic (68). Surprisingly, this strategy did not demonstrate a significant impact in reducing the risk of overdose (68). This outcome points to the complexities involved in managing the drug toxicity crisis and underscores the necessity for multifaceted and nuanced strategies.

The study emphasizes the dynamic nature of opioid misuse risk factors, highlighting the need for ongoing research and adaptation to the shifting landscape of drug use and public health issues. Utilizing the BC Provincial Overdose Cohort and its administrative data, this research project will enable a detailed examination of various factors and their relationships with both general and fatal overdoses. This approach is vital for developing more effective strategies to address the complexities of the drug toxicity crisis.
1.6.4 The Usage of the BC Provincial Overdose Cohort

The expansive and intricate nature of the BC Overdose Cohort database makes it an ideal candidate for employing Machine Learning (ML) techniques to discern and forecast individual risk factors for overdose. The database's assembly of various risk and protective factors offers a unique chance to apply ML algorithms. These algorithms are adept at modeling complex, non-linear relationships between independent and dependent variables, such as a history of injecting substances and fatal overdose. ML algorithms can construct diverse models using nodes that represent independent features within each layer of a mathematical equation, thereby creating a nuanced network of relationships with varying weights. Deep learning, a subset of ML, further enhances this by introducing multiple layers and feedback loops with feed-forward mechanisms. These feedback loops are particularly beneficial, allowing for continual refinement of the algorithm based on real-world performance. Utilizing statistical and mathematical methodologies, ML also enables the application of advanced programming languages like R and Python for data manipulation and analysis. This synergy between data science and computer science magnifies the potential of these programming languages for intricate data wrangling tasks.

Building on this, the BC Provincial Overdose Cohort is suitable to serve as a foundation for developing a predictive model for fatal and general overdose. Our Addiction and Concurrent Disorders (ACD) research team plans to harness this robust provincial database to pinpoint key risk factors for overdose. This model will form the cornerstone of a Clinical Decision Support System (CDSS), the existing Risk Assessment and Management Platform (RAMP), enhancing its accuracy at predicting risk, and increasing its efficacy in real-world applications.
1.7 Machine learning

Machine Learning (ML) is a branch of Artificial Intelligence (AI), which in turn falls under the broader category of Data Science. ML draws on a variety of disciplines, including Mathematics, Statistics, Computer Science, and Data Analysis. It encompasses a diverse array of techniques and methodologies, each tailored to the specific goals of a project. Generally, there are two practical approaches to ML: Supervised Learning and Unsupervised Learning. In supervised ML, specifically predictive modelling, scientists define the specific features of a problem and feed these features/characteristics/variables into a machine (69). These include dependent and independent features, in which the machine will try to learn from the pre-existing data to predict future outcomes (69). This method then divides a labeled data set into two parts consisting of training and testing data. It uses the training data to fit the model to learn how the independent and dependent features are correlated as well as use the testing data to validate it.

Contrastingly, unsupervised learning is more exploratory in nature, focusing on inference in datasets lacking pre-defined labels. It seeks to uncover hidden structures and interrelationships among features. However, given the aim of this project to develop a predictive instrument, our focus will be primarily on supervised learning techniques. These methods are notably effective in tackling complex datasets, which are often characterized by poor quality, including inaccuracies, missing elements, or mismatches. Supervised learning techniques have successfully been utilized in healthcare settings to predict the probability of correct diagnoses, such as for diabetes or mental health disorders, based on extracted data from medical records (70,71). The results in highly accurate models demonstrate that these systems can accurately predict the risk of adverse outcomes for both clients and practitioners (72). One also could use natural language processing
approaches to analyze psychiatrists’ notes for the prediction of suicidal behaviors (73). This project aims to utilize existing data to create a model capable of assisting decision-making in a clinical setting to address the opioid overdose crisis. It will aspire towards leaving a profound impact at the micro or individual level. ML is a perfect candidate in terms of the feasibility in utilizing predictive modelling to create a tool appropriate for addressing individual risk behavior. This tool can aid towards the risk assessment and monitoring of behaviors that will play a crucial role in individualized addiction treatment and next-generation personalized medicine (73).

The foundation of ML lies in its reliance on statistical, linear algebraic, probabilistic, and calculus principles. These mathematical underpinnings are critical for modifying model parameters and for the logical reasoning behind the model's structure and function. In the context of this project, ML's suitability is evident in its capacity to process and utilize large-scale data for aiding clinical decision-making, particularly in addressing the opioid overdose crisis. ML's advantage over traditional statistical models is its emphasis on predictive accuracy rather than solely on understanding relationships between variables. ML models vary in interpretability, from the highly interpretable, such as lasso regression, to the more opaque, like neural networks within deep learning.

A vital aspect of ML is its application across interdisciplinary fields, facilitated by advanced computing languages like R and Python. These languages provide the tools for seamless data manipulation, leveraging developments in data and computer science. For example, RStudio packages such as 'tidy' and 'dplyr' are instrumental in the design and development of derived features in ML models, and tools like R plumber enable the integration of these models into cloud-based system application programming interfaces (APIs). This
integration allows platforms like RAMP to harness the predictive capabilities of the models, addressing individual risk factors.

Moreover, the cloud-based deployment of this project's ML model introduces an innovative dimension to its application - the creation of a self-educating system through an automated pipeline. The model not only tracks predictive values but could also integrates these new data points back into the training process. This continual loop of feedback and learning allows the model to evolve, refining its predictions and adapting to new and emerging data trends. Such an approach exemplifies the adaptive and progressive nature of ML, highlighting its potential to revolutionize predictive modeling and decision-making processes in various sectors, especially in healthcare. This evolving model dynamically adjusts its focus on different features based on real-world data developments, showcasing the transformative potential of ML in research and practical applications, particularly in personalized medicine and individualized treatment strategies.

1.7.1 Machine learning and statistics

One crucial consideration is the choice to employ Machine Learning (ML) methods over other statistical and mathematical techniques, such as additive/spline-based regression models, mixed-effects/hierarchical models, bootstrapping, and permutation tests. These traditional methods are undoubtedly effective in discovering correlations between dependent and independent features. However, the decision to use ML in our project is driven by its specific aims and the nature of the data at hand. Our project's strength lies in leveraging the vast amounts of existing big data to develop a model that aids in clinical decision-making, particularly in the context of the opioid overdose crisis. The goal is to make a tangible impact at the individual
level, and ML is uniquely suited for this task. Its predictive modeling capabilities allow us to construct a tool specifically designed to address individual risky behaviors. This approach, rooted in the practical application of ML, aligns closely with the project's objective of employing data-driven insights to mitigate individual risks and contribute meaningfully to the healthcare domain.

The fundamental differences between statistics and machine learning lie not only in their objectives but also in their methodologies, techniques, and underlying goals. Statistics, a discipline with deep roots in mathematics, is primarily concerned with making inferences from data. It focuses on understanding relationships, testing hypotheses, and estimating parameters. Statistical methods often begin with a model based on assumptions about the nature of data, such as its distribution, variance, and other properties. The statistical approach is theory-driven; it seeks to explain or describe phenomena through careful analysis and interpretation of data. This involves choosing models that incorporate system knowledge or dataset characteristics, such as linear regression models or ANOVA for testing differences between groups. Statistical analysis is deeply concerned with the validity, assumptions, and interpretability of these models.

Machine Learning (ML), on the other hand, is a field rooted in computer science and is predominantly focused on the development and application of algorithms that can learn from and make predictions or decisions based on data. Unlike traditional statistical methods, ML does not necessarily begin with a predefined hypothesis or model. Instead, it often employs data-driven techniques to develop predictive models, where the primary goal is not to test a hypothesis but to achieve the highest possible accuracy in prediction. ML utilizes a wide array of algorithms, from simpler models like decision trees to more complex ones like neural networks and support vector
machines. These models, particularly in deep learning, can handle vast and complex datasets but are often considered "black boxes" due to their lack of interpretability.

Moreover, the methodologies in ML are characterized by their emphasis on performance metrics such as accuracy, precision, recall, and the ROC curve. The process often involves splitting data into training and testing sets to evaluate the model's performance and its ability to generalize to new, unseen data. This focus on empirical results and prediction accuracy is what distinctly sets ML apart from traditional statistical methods.

Another key difference is in how each field handles data. Statistics typically requires careful data selection and rigorous assumptions about data quality and structure. In contrast, ML can often work with and even benefit from large, unstructured, or semi-structured datasets, applying techniques like feature extraction and dimensionality reduction to make sense of this data.

In summary, while both statistics and machine learning aim to draw insights from data, they diverge in their approaches, techniques, and goals. Statistics is more concerned with understanding and explaining data based on theoretical foundations, whereas machine learning emphasizes predictive accuracy and decision-making, often leveraging complex algorithms and large datasets. This distinction reflects a shift from a traditional, theory-driven approach in statistics to a more empirical, performance-driven approach in machine learning.

### 1.7.2 Supervised machine learning problem

In the realm of supervised machine learning, regression and classification stand as the two primary types of problems, each defined by unique goals and methodologies (74). Regression, at its core, is aimed at predicting continuous outcomes and involves establishing a
relationship between a dependent variable and one or more independent variables. The spectrum of regression analyses is diverse, encompassing methods like linear regression, which forecasts an outcome as a linear function of input features, and polynomial regression, which models the relationships as polynomials to accommodate the intricacies of non-linear data. Furthermore, techniques such as Ridge and Lasso regression bring in the concept of regularization, effectively preventing overfitting by penalizing large coefficients, thus maintaining the model's generalizability (75).

On the other hand, classification focuses on predicting discrete labels, essentially categorizing data into pre-established classes. The methodologies in classification are varied, ranging from simpler, more interpretable models like decision trees, which segment the data space into subsets based on feature values, to more complex and computationally intensive models like neural networks. These neural networks, especially in deep learning, are capable of capturing complex patterns in large datasets but often at the cost of reduced interpretability, known as the 'black box' issue.

Moreover, machine learning extends these concepts with advanced techniques. For instance, ensemble methods like Random Forests combine multiple decision trees to improve predictive performance and robustness (76). Support Vector Machines (SVMs) are another powerful tool, particularly effective in high-dimensional spaces, which find the optimal hyperplane for classification tasks (77).

These two paradigms - regression and classification - while distinct in their application, share a common ground in machine learning: both require careful consideration of feature selection, model tuning, and validation. The choice between regression and classification in a machine learning project hinges on the nature of the target variable - continuous for regression
and categorical for classification. This decision significantly influences the model selection, algorithm implementation, and ultimately, the interpretation of the results. The intricacy of machine learning models, be it in regression or classification, underscores the importance of understanding the underlying data patterns, model assumptions, and the specific problem context to effectively harness these powerful tools in predictive analytics.

In this project, our focus will be on employing classification models since our primary outcome, overdose, is categorical in nature. As detailed in chapter 2, many issues in addiction psychiatry, particularly those pertaining to illicit drug use, hinge on predicting categorical outcomes, which are often binary variables.

### 1.7.3 Machine learning in mental health

Machine learning (ML) has recently become a focal point in mental healthcare (78). Its impact spans several crucial areas, notably in enhancing the prediction of diseases and improving diagnostic accuracy. These advancements in early detection and accurate diagnosis are vital for effective treatment and management of mental health conditions (79).

In addition, ML plays a significant role in personalizing treatment plans. By analyzing large datasets, ML algorithms can identify patterns and factors that influence the effectiveness of different treatments, leading to more customized and effective care for patients. This personalization is particularly crucial in mental health, where individual patient responses to treatments can vary widely (80,81). However, it's important to note that while ML offers these promising avenues, some applications have yielded underwhelming results. The continued exploration and refinement of ML techniques in mental healthcare are essential to realize their full potential. Furthermore, ML contributes to predicting patient outcomes. By processing and
learning from historical data, ML models can forecast potential future mental health crises or relapse, enabling pre-emptive care and interventions. This predictive capability is instrumental in proactive mental healthcare management (82). ML also aids in data collection and monitoring. It automates the gathering and analysis of patient data, providing healthcare professionals with insights that inform treatment decisions and patient care strategies (83). Another significant aspect of ML in mental healthcare is its potential to address treatment accessibility issues, particularly for vulnerable and marginalized populations in rural areas (84). ML can offer cost-effective, accessible solutions, bridging the gap in mental health services for these communities (84). This is especially important in areas where resources and specialized care are limited, demonstrating the far-reaching implications of ML in democratizing mental healthcare (84). Chapter 2 will highlight machine learning models that pertain to illicit opioid use in addiction psychiatry.

1.7.4 Project Aim

The aim of this project is to develop a machine learning solution that can overcome barriers and enhance capacity. Our goal is to empower individuals and non-healthcare specialists with the tools to assess overdose risk. This solution takes the shape of a predictive model within a clinical decision support system (RAMP) designed to identify key risk factors, thereby enabling individuals to take control of their situation through informed understanding. This approach not only fosters self-awareness but also democratizes access to crucial health information, paving the way for proactive personal health management.

The objective of this project is to create a ML predictive tool that predicts the risk of overdose and general overdose.
1.8 Research Question

Can machine learning methods be utilized to predict overdose?

1.9 An Overview of the following chapters

The subsequent chapters delineate the work conducted during this PhD research project. Chapters 2 through 6, along with Appendices A, B, and C, encompass work that has either been published or is presently under review for publication.

Chapter 2 presents a systematic review on the application of machine learning in predicting opioid overdose outcomes. The objective of this study is to conduct a thorough exploration of existing research in the field, examining evidence that either supports or refutes the hypothesis that machine learning can effectively assess the risk of both fatal and general opioid overdoses. This involves a critical review of the methodologies used in training, testing, and evaluating machine learning models in this context, and an assessment of their ability to predict various opioid overdose outcomes. A meta-analysis will be performed to analyze the collective performance of these models. Appendix A will provide additional material gleaned from this comprehensive review.

Chapter 3 presents a systematic review focuses on the use of Clinical Decision Support Systems (CDSS) in addiction and concurrent disorders, examining evidence that either supports or refutes the hypothesis regarding the effectiveness of CDSS in enhancing clinical decision-making processes. A meta-analysis is conducted to compare the effect sizes from the implementation of CDSS against standard treatment practices.
Chapter 4 delves into the exploratory data analysis, data wrangling, and preprocessing undertaken with the BC Provincial Overdose Cohort data, aiming to predict overdose outcomes. This chapter also includes trials and tribulations related to preliminary models.

Chapter 5 outlines the results of the machine learning methodology and discusses the implications of the models developed for predicting fatal and general overdoses.

Chapter 6 illustrates the exploratory data analysis conducted to establish explanatory models and survival analyses.

Finally, Chapter 7 synthesizes the findings of Chapters 1 through 6 and Appendices A to C, encapsulating the results and conclusions. It also addresses the significance and limitations of this research, as well as envisaged future steps.
Chapter 2: Utilizing machine learning for early intervention and risk management in the opioid overdose crisis: a systematic review and meta-analysis in addiction psychiatry

2.1 Introduction

The mental health crisis, particularly the surge in opioid overdoses and resulting fatalities, has emerged as a significant concern. In North America, the crisis is exacerbated by limited healthcare access due to insufficient resources and expertise (85,86). The U.S. and Canada have seen overdose rates triple, leading to a decrease in life expectancy (87). BC responded to the rise in opioid-related deaths by declaring a public health emergency in 2016. By 2022, the annual death toll had climbed to 2377, and in 2023, it reached 2511 deaths per year (19).

The mental health crisis, particularly the opioid epidemic, necessitates innovative healthcare solutions. The emergence of big data offers a chance to apply evidence-based approaches to this complex issue. Machine Learning (ML) stands out as a transformative technology for healthcare, enabling early interventions. The need for a paradigm shift in healthcare delivery is critical, given the more than tripling rates of overdose harm and deaths in North America (88–90). This review examines ML's role in tackling the drug toxicity crisis, highlighting the urgent need for change in healthcare practices. ML's proficiency in handling vast, intricate datasets, combined with its predictive accuracy, is crucial for identifying patients at risk and foreseeing overdose occurrences for prompt clinical responses.

Traditional predictive modeling, often based on regression analysis known for its interpretability, contrasts with Machine Learning's (ML) empirical strength in managing complex data (91). The main difference lies in ML's focus on predictive accuracy, sometimes
Deep Learning (DL), a subset of ML, processes large volumes of data using multi-layer neural networks, offering a nuanced analysis of complex patterns. This positions ML and DL as key elements in developing mental healthcare, particularly in addiction psychiatry, by providing a deeper understanding of data intricacies.

In the medical field, the recent surge in Machine Learning (ML) applications has been notable. Its deployment spans from diagnosing diverse disorders, including neurocognitive diseases, depression, and anxiety (92–94). Within addiction psychiatry, supervised ML techniques are now frequently used for categorizing current drug users, evaluating future abuse risks, and forecasting treatment results (95,96). ML models play a pivotal role at the system level, offering insights to clinicians and patients by pinpointing key factors that contribute to adverse events, potentially increasing mortality risk and adversely affecting individuals' lives. Risk behaviors such as solitary drug use, polysubstance abuse, and intravenous drug administration are significant contributors to severe outcomes, including overdoses and mental health crises. The overarching aim of ML models in this context is to identify, reduce, and prevent such occurrences by impacting high-risk behaviors.

In scenarios like high-risk substance use, various risk factors, both changeable and unchangeable, contribute to critical outcomes like overdoses (97). Understanding these factors is essential due to the growing overdose crisis, as it can help change the course of substance use disorders and prevent serious events or deaths (98–100). Identifying key risk factors for overdose and targeting prevention strategies is crucial. While the use of Machine Learning (ML) in psychiatry is expanding, its application in addiction psychiatry needs more in-depth exploration (101–103). ML tools assist in reducing risks, promoting early interventions, and preventing deaths at the individual level. Clinically, they are valuable for monitoring and tracking patients'
progress in addiction treatment. Research that compares various ML models with clinical
diagnostic tools offers insights into diverse risk factors.

This systematic review and meta-analysis aims to achieve two primary objectives: firstly, to scrutinize the current literature utilizing Machine Learning (ML) models in the classification of outcomes associated with illicit opioid use; and secondly, to assess the efficacy of these ML models, examining their types and the accuracy of the outcomes they forecast.

2.2 Methods

2.2.1 Protocol

The study's protocol was developed in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (104) and was further refined through consultations with the Cochrane Handbook for Systematic Reviews and Interventions (105). The protocol underwent registration in Prospero (106). In accordance with the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans, seeking approval from the ethics committee was deemed unnecessary for this study (107).

2.2.2 Search Strategy

A detailed search strategy for this review was crafted by a medical research librarian (VK) in collaboration with experts (AT) for Ovid MEDLINE. This strategy was then adapted for other databases, using comparable subject headings where needed. The strategy was applied to databases like Embase, CINAHL, PsycINFO, and Web of Science, with institutional access through the University of British Columbia. The search methodologies for all these databases are
detailed in Appendix A.2. The review encompasses studies that met the inclusion criteria from searches up to July 19, 2023.

2.2.3 Eligibility Criteria

Data collection for the study was performed by two independent reviewers (AT, JK). They gathered information including study demographics, introduced features, significant features, evaluation metrics, and predictive outcomes. Studies lacking explicit Area Under the Receiver Operating Characteristic curve (AUC/AUROC) metrics were excluded from the meta-analysis but included in qualitative data analysis. The data collection template was designed after a thorough review of the manuscripts. Bias risk evaluation across all studies was conducted by (RM) using the RoB 2 tool for randomized trials by Cochrane, the Modified Newcastle Ottawa Assessment Scale for various study types, and the CASP tool for qualitative assessments (108,109).

2.2.4 Risk of bias

Two reviewers, (AT, RM) conducted the risk of bias studies independently. Bias risk evaluation across all studies was conducted by using the Modified Newcastle Ottawa Assessment Scale for various study types. In cases of disagreement on assessed items, resolution was sought by consulting a third reviewer (JK) for their opinion.

2.2.5 Data Extraction

Data collection for the study was performed by two independent reviewers (AT, JK). They gathered information including study demographics, introduced features, significant
features, evaluation metrics, and predictive outcomes. Studies lacking explicit Area Under the Receiver Operating Characteristic curve (AUC/AUROC) metrics were excluded from the meta-analysis but included in qualitative data analysis. The data collection template was designed after a thorough review of the manuscripts.

### 2.2.6 Methodology for Meta-Analysis

AUC was the chosen metric for comparing model performance. This metric is optimal for model evaluation, balancing sensitivity and specificity and offering robust performance in imbalanced datasets. AUC provides a non-parametric, interpretable measure ranging from 0.5 (no discriminative ability) to 1 (perfect discrimination), making it a comprehensive and versatile metric for assessing model performance in classification tasks (110,111). Meta-analyses used standardized mean difference transformed AUC estimates. AUC were transformed using standard formula (112), their variance was estimated using modified Wald statistics with continuity correction (113). Due to multiple AUC reports for various models and outcomes, a three-level meta-analysis was employed. Subgroups were analyzed based on model and outcome types, using Standardized Mean Differences (SMD) and implemented in R software, Version 4.3.2 (114) using the package metafor V. 4.4-0 (115).
2.3 Results

Figure 8 ML PRISMA flow diagram (104)
### Table 1: Qualitative Analysis of studies for Addiction & Concur Disorders: Systematic Review (Short, full version can be viewed in Appendix A)

<table>
<thead>
<tr>
<th>Study</th>
<th>Demographic Population</th>
<th>Machine learning models and algorithms</th>
<th>Evaluation Metric</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>(116)</td>
<td>Substance use population, Age, Gender, Race</td>
<td>Supervised Classification: Logistic Regression, Random Forest</td>
<td>Accuracy, AUROC, Sensitivity</td>
<td>Predicting adults at risk for OUD</td>
</tr>
<tr>
<td>(117)</td>
<td>2016 edition of the National Survey on Drug Use and Health (NSDUH)</td>
<td>Random forest classifier</td>
<td>AUROC, Sensitivity</td>
<td>Predicting risk for OUD</td>
</tr>
<tr>
<td>(118)</td>
<td>Hospital and health care population, Age, Sex, Gender</td>
<td>Supervised Deep learning: Convolutional neural network</td>
<td>AUROC, Negative predictive value</td>
<td>Validating opiate misuse in hospital</td>
</tr>
<tr>
<td>(119)</td>
<td>Substance use population, Age, Education, U.S. Family Income</td>
<td>Supervised Regression: Elastic net</td>
<td>AUROC</td>
<td>Identifying markers for opiate and alcohol dependence</td>
</tr>
<tr>
<td>(120)</td>
<td>1 year of S-11 emergency medical CAD and ePCR (including national administration data)</td>
<td>Random Forest</td>
<td>AUROC</td>
<td>General opioid overdose</td>
</tr>
<tr>
<td>(121)</td>
<td>4,681,809 unique veterans, 616,827 of whom had a hospitalization</td>
<td>Naive Bayes model</td>
<td>AUROC</td>
<td>Substance abuse</td>
</tr>
<tr>
<td>(122)</td>
<td>Iranian EHR opioid users</td>
<td>Decision tree and multiple logistic regression (MUR)</td>
<td>Sensitivity, Specificity, Accuracy</td>
<td>Identifying people at risk for drug abuse</td>
</tr>
<tr>
<td>(123)</td>
<td>Emergency Department Visits, over 12 years old between Jan 2016 and March 2018</td>
<td>Logistic regression</td>
<td>F1, AUROC</td>
<td>Diagnosis of OUD</td>
</tr>
<tr>
<td>(124)</td>
<td>Hospital and health care population, Age, Sex, Hospital and health care population</td>
<td>Supervised Deep learning: Convolutional neural network</td>
<td>AUROC, Sensitivity</td>
<td>Identifying opiate misuse</td>
</tr>
<tr>
<td>(125)</td>
<td>Traditional Clinic population</td>
<td>Random forest</td>
<td>N/A</td>
<td>Security of opioid overdose events</td>
</tr>
<tr>
<td>(126)</td>
<td>Administrative EHR data</td>
<td>Logit Regression, Random Forest, Deep Learning Network</td>
<td>AUROC, Precision, F1, Recall</td>
<td>Predicting OUD</td>
</tr>
<tr>
<td>(127)</td>
<td>Patients receiving medication for opioid use disorder</td>
<td>Logistic regression, logit ridge regression, XGBoost</td>
<td>AUROC (95% CI), AUPR, Sensitivity, Specificity, PPV, NPV</td>
<td>DUG Treatment Advisement</td>
</tr>
<tr>
<td>(128)</td>
<td>National Survey on Drug Use and Health</td>
<td>Random forest</td>
<td>N/A</td>
<td>Classify heroin use</td>
</tr>
<tr>
<td>(129)</td>
<td>Patients receiving medication for opioid use disorder</td>
<td>Support Vector Machines</td>
<td>Accuracy, Sensitivity, PPV, NPV</td>
<td>Self-reported opioid craving</td>
</tr>
<tr>
<td>(130)</td>
<td>Substance use population, Age, Gender, Race, Age, Medicare, Medicaid</td>
<td>Supervised Regression: Stepwise Regression</td>
<td>Accuracy, AUROC, Sensitivity, Specificity</td>
<td>Predicting Future Chronic Opioid Use Among Hospitalized Patients</td>
</tr>
<tr>
<td>(131)</td>
<td>Health record data from 2011 to 2022</td>
<td>Logistic regression</td>
<td>N/A</td>
<td>Risk for future chronic opioid therapy</td>
</tr>
<tr>
<td>(132)</td>
<td>1,014,333 Maryland residents aged 18-80 with at least 1 opioid prescription</td>
<td>Logistic regression</td>
<td>N/A</td>
<td>Risk for future opioid overdose</td>
</tr>
<tr>
<td>(133)</td>
<td>Hospital and health care population</td>
<td>Regression: logistic regression</td>
<td>N/A</td>
<td>Predicting Opioid Use Disorder</td>
</tr>
<tr>
<td>(134)</td>
<td>Hospital and health care population, Substance use population, Age, Sex, Race, Mortality</td>
<td>Supervised Classification: Support vector machine, Random forest, Recurrent neural network, Linear regression</td>
<td>AUROC, K Cohen’s Kappa, Sensitivity</td>
<td>Predicting opioid dependence and long-term opioid use</td>
</tr>
<tr>
<td>(135)</td>
<td>Hospital and health care population, Substance use population, age</td>
<td>Supervised Classification: Decision tree, Deep learning: Multi-layer perceptron model, Logistic regression</td>
<td>AUROC, Sensitivity</td>
<td>Risk of Drug Intoxication Mortality</td>
</tr>
<tr>
<td>(136)</td>
<td>Substance use population, Age, Sex, Region of USA</td>
<td>Regression: Multivariate logistic regression</td>
<td>AUROC, Sensitivity</td>
<td>Opioid Overdose Prediction</td>
</tr>
<tr>
<td>(137)</td>
<td>Substance use population, Age, Sex, number of family, household income</td>
<td>Supervised Classification: Support vector machine, Multinomial logistic regression</td>
<td>AUROC, Sensitivity</td>
<td>Opioid Overdose Prediction</td>
</tr>
<tr>
<td>(138)</td>
<td>Hospital and health care population, Substance use population, Age, Gender, Race</td>
<td>Supervised Classification: Random Forest, Decision Tree, Deep learning: Deep learning</td>
<td>Accuracy, AUROC, F1</td>
<td>Prediction of Opioid Overdose Risk</td>
</tr>
<tr>
<td>(139)</td>
<td>Common’s Health Facts database with over 3 million patients</td>
<td>Random Forest, Decision Tree, Logit Regression: CRNN, LSTM, B-LSTM, Attention, Transformer: UPTED LSTM, Fatced-LSTM</td>
<td>Precision, Recall, F1, AUROC</td>
<td>Opioid Overdose Risk</td>
</tr>
<tr>
<td>(140)</td>
<td>United States</td>
<td>Supervised Classification: Random Forest, Decision Tree, Deep learning: Long Short term memory, Dense neural network, Logistic regression</td>
<td>AUROC, F1</td>
<td>Predicting opioid overdose risk in patients with opioid prescriptions</td>
</tr>
<tr>
<td>(141)</td>
<td>Electronic health records of patients who have been prescribed with medications containing active opioid ingredient</td>
<td>Supervised Classification: Random forest, Decision tree, Deep learning: Long short term memory, Dense neural network, Logistic regression</td>
<td>AUROC, F1, Precision</td>
<td>Risk of opioid use disorder</td>
</tr>
<tr>
<td>Study</td>
<td>Demographic Population</td>
<td>Machine learning models and algorithms</td>
<td>Evaluation Metric</td>
<td>Prediction</td>
</tr>
<tr>
<td>-------</td>
<td>------------------------</td>
<td>----------------------------------------</td>
<td>------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>142</td>
<td>United States Hospital and health care population Substance use population, Age, Sex, Gender United States Hospital and health care population Substance use population</td>
<td>Supervised Classification: Random Forest</td>
<td>AUC</td>
<td>Predicting opioid dependence</td>
</tr>
<tr>
<td>143</td>
<td>Maryland residents aged 18-80 years with a filled opioid prescription (n=565,175) from January to June 2016</td>
<td>Logistic regression</td>
<td>AUC, Sensitivity, Specificity, PPV, NPV</td>
<td>opiate-related overdose death</td>
</tr>
<tr>
<td>144</td>
<td>Maryland residents aged 18 to 80 years with an opioid fill between April to June 2013</td>
<td>Logistic regression</td>
<td>AUC, Sensitivity, Specificity, PPV, NPV</td>
<td>opiate-related overdose death</td>
</tr>
<tr>
<td>145</td>
<td>414,259 patients; 260,745 were females, claims data from 2009 to 2014; Medicaid patients; Deep learning models include long short-term memory model, transformer model, and a novel deep learning model</td>
<td>Logistic regression, random forest, and ensemble methods are used</td>
<td>Accuracy, Precision, Recall</td>
<td>Predicting DOJ</td>
</tr>
<tr>
<td>146</td>
<td>Hospital and health care population Substance use population Net Income</td>
<td>Supervised Regression: Linear regression, Lasso regression; Ridge regression, Elastic net</td>
<td>Net Returns</td>
<td>Predicting hospital Opioid admissions</td>
</tr>
<tr>
<td>147</td>
<td>Pennsylvania residents with a prescription dispensed from February 2016 to September 2021, 474,208 patients; 269,748 were females.</td>
<td>Logistic regression, random forest, neural network, gradient boosting machine, and a super learner (ensemble learning model)</td>
<td>AUC, C-Statistics</td>
<td>Predicting total overdose</td>
</tr>
<tr>
<td>148</td>
<td>40,630 patients taking chronic opioid therapy and externally validated the model in 10,703 patients</td>
<td>Cox proportional hazards regression</td>
<td>C-Statistics</td>
<td>Predicting 3-Year Overdose Risk</td>
</tr>
<tr>
<td>149</td>
<td>985 males with substance use disorder</td>
<td>Gradient boosting</td>
<td>Importance</td>
<td>substance craving</td>
</tr>
<tr>
<td>150</td>
<td>705 program participants, Houston Emergency Opioid Engagement System (HEROS)</td>
<td>Random forest</td>
<td>Sensitivity, specificity</td>
<td>Adherence to program</td>
</tr>
<tr>
<td>151</td>
<td>9680 Pennsylvania Medicaid beneficiaries</td>
<td>XGBoost C-Statistics, Sensitivity, Specificity, PPV, NPV, AVE</td>
<td>predicts risk of death after a non-fatal opioid overdose</td>
<td></td>
</tr>
<tr>
<td>152</td>
<td>Youth population, Age, Education, ID, Gender, Race</td>
<td>Supervised Classification: distributed random forest, and gradient boosting machine Deep learning: artificial neural network</td>
<td>Accuracy, AUC, AUROC</td>
<td>Risk for Problem Opioid Use</td>
</tr>
<tr>
<td>153</td>
<td>5196 patients who were commercially insured, initial supervised learning between January and December 2016</td>
<td>Logistic regression, decision tree, random forest, extreme gradient boosting, support vector machine, and artificial neural network</td>
<td>Precision, Recall, F1 score, C-statistic</td>
<td>Adherence to program</td>
</tr>
<tr>
<td>155</td>
<td>Hospital and health care population, Age, Gender, Ethnicity Hospital and health care population</td>
<td>Supervised Regression: logistic regression model</td>
<td>AUC, Sensitivity, Specificity, Other Measurements</td>
<td>Risk for Problem Opioid Use</td>
</tr>
<tr>
<td>156</td>
<td>200000 patients aged 18-80 years with a filled opioid prescription (n=565,175)</td>
<td>Logistic regression, Deep learning</td>
<td>accuracy, precision, recall, sensitivity, F1 score, specificity, and AUC-ROC</td>
<td>Opioid Prescription and OUD</td>
</tr>
<tr>
<td>157</td>
<td>10 years or older who filed an opioid prescription between 2014 and 2019 and who were opioid-naive, insurance coverage</td>
<td>Supervised Machine learning: ensemble machine-learning model</td>
<td>Accuracy, Sensitivity, Specificity, AUC</td>
<td>Opioid use disorder</td>
</tr>
<tr>
<td>158</td>
<td>Substance use population, Age, Sex, Race, Ethnicity, Disability as the reason for Medicare eligibility, Receipt of town income subsidy, and urbaniy of county of residence</td>
<td>Classification: random forest, gradient boosting machine Deep learning: artificial neural network</td>
<td>Multinomial logistic regression, most minor absolute shrinkage, and selection operator-type regression</td>
<td>Risk for fatal overdose</td>
</tr>
<tr>
<td>159</td>
<td>United States Substance use population, Sex, Race, Class, Overall health United States Substance use population</td>
<td>Supervised Regression: elastic net</td>
<td>AUC, C-Statistics, Sensitivity, Specificity, Other Measurements</td>
<td>Predicting risk of incident opioid use disorder</td>
</tr>
<tr>
<td>160</td>
<td>Hospital and health care population, Age, Sex, Race, Type of Medical specialty, and Duration of continuous enrollment in Medicaid each month Hospitals and health care population</td>
<td>Supervised Classification: Gradient Boosting Model</td>
<td>AUC</td>
<td>Predicting risk of opioid overuse among Medicaid beneficiaries</td>
</tr>
<tr>
<td>161</td>
<td>administrative claims data of Medicaid beneficiaries in Pennsylvania from Jan 1, 2013, to Dec 31, 2016</td>
<td>Supervised Classification: Gradient Boosting Model</td>
<td>AUC</td>
<td>Predicting 3-month risk of opioid overdose</td>
</tr>
<tr>
<td>162</td>
<td>Patients from several primary care clinics across the United States</td>
<td>Supervised Gradient boosting</td>
<td>Sensitivity, Specificity, Recall</td>
<td>Opioid use disorder</td>
</tr>
<tr>
<td>163</td>
<td>Hospital data obtained from a Level I Trauma Center Hospital in Wisconsin</td>
<td>Logistic regression, random forest, support vector machine, AdaBoost, XGBoost</td>
<td>accuracy, precision, recall, and AUC</td>
<td>Opioid Overdose and 90 day all-cause readmission</td>
</tr>
<tr>
<td>164</td>
<td>Substance use population, Race, Age, Gender, Country of residence, and level of education</td>
<td>Supervised Deep learning: artificial neural network (ANN) Deep learning: artificial neural network (ANN)</td>
<td>accuracy, precision, recall, and AUC</td>
<td>Classifying a drug user and predicting last drug use</td>
</tr>
<tr>
<td>165</td>
<td>308 individuals with BD from a specialized unit</td>
<td>Random Forest, Multiple logistic regression</td>
<td>Accuracy, Sensitivity, Specificity, F1-score</td>
<td>SUD</td>
</tr>
<tr>
<td>166</td>
<td>survey data from 37 schools, 6592 youth in grades 7-12</td>
<td>Recursive partitioning, a Type of decision tree-based Machine Learning</td>
<td>Accuracy, Sensitivity, Specificity</td>
<td>opioid use in the last 30 days</td>
</tr>
<tr>
<td>167</td>
<td>590 young adults aged 18-24 years from Patras, Greece</td>
<td>Logistic regression</td>
<td>AUC, specificity, sensitivity, ROC</td>
<td>illicit substance abuse or dependence</td>
</tr>
<tr>
<td>168</td>
<td>Multinomial Logistic Regression (MLR), Support Vector Machines (SVM), Decision trees (DT), Random Forest (RF), Gradient Boosting Decision trees (GBDT)</td>
<td>Accuracy, F1, Cohen’s kappa, ROC</td>
<td>Risk of alcohol and drug misuse</td>
<td></td>
</tr>
<tr>
<td>169</td>
<td>Youth population, Age, Sex, State of residence (USA)</td>
<td>Supervised Logistic regression</td>
<td>OR</td>
<td>Risk of opioid abuse</td>
</tr>
<tr>
<td>170</td>
<td>Maryland residents with 1 or more records in 2015 in any of 4 available data sets (PDMP prescriptions, statewide hospital inpatient and emergency department visits, public-sector specialty behavioral health care admissions, or sheet, prescriptions)</td>
<td>Logistic regression</td>
<td>AUC, sensitivity, specificity, PPV, NPV</td>
<td>Fatal and general overview</td>
</tr>
<tr>
<td>Study</td>
<td>Demographic Population</td>
<td>Machine learning models and algorithms</td>
<td>Evaluation Metric</td>
<td>Prediction</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------</td>
<td>------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>(171)</td>
<td>All accidental opioid-involved overdose deaths in Rhode Island between July 1, 2010, and June 30, 2010</td>
<td>LASSO, random forest</td>
<td>Accuracy</td>
<td>Opioid overdose death rate</td>
</tr>
<tr>
<td>(172)</td>
<td>Hospital and health care population: Substance use population, Sex, Age, Hospital and health care population, Substance use population</td>
<td>Supervised Classification: Gradient boosting, Deep learning, Random Forest</td>
<td>AUROC</td>
<td>Detection of opioid use disorder</td>
</tr>
<tr>
<td>(173)</td>
<td>Patients with at least one opioid prescription within a U.S. commercial claims database (2001-2015)</td>
<td>Logistic regression</td>
<td>C-statistic, sensitivity, specificity, PPV, NPV</td>
<td>Opioid overdose</td>
</tr>
<tr>
<td>(174)</td>
<td>1,468 homeless youth across six states in USA, namely California (CA), Arizona (AZ), Colorado (CO), Missouri (MO), Texas (TX), and New York (NY), from June 2016 until July 2017</td>
<td>Logistic regression, classification and regression tree (CART), conditional inference forest (CForest), adaptive boosting (AdaBoost), extreme gradient boosting (XGBoost), support vector machine (SVM), multi-layer perceptron (MLP)</td>
<td>Accuracy, precision, recall, F1, AUROC</td>
<td>Susceptibility to SUD</td>
</tr>
<tr>
<td>(175)</td>
<td>Random individuals whose first opioid prescriptions were made between 01/01/2012 and 03/01/2016</td>
<td>Logistic regression</td>
<td>Accuracy</td>
<td>Risk factors for OUD</td>
</tr>
<tr>
<td>(176)</td>
<td>Opioid-naïve patients aged 12 years and above prescribed an opioid analgesic in California between 2016 and 2019</td>
<td>Logistic regression, extreme gradient boosting (XGBoost), Linear regression</td>
<td>c-statistic, calibration slope, intercept, sensitivity, specificity</td>
<td>Transition to long-term (&gt;90 d) opioid use</td>
</tr>
<tr>
<td>(177)</td>
<td>Adolescents aged between 14-18 years enrolled in a Colorado healthcare system</td>
<td>Care proportional hazards regression analysis, Random forest, Support vector machine, Neural network, Decision tree, logistic regression</td>
<td>ROC</td>
<td>Opioid use disorder</td>
</tr>
<tr>
<td>(178)</td>
<td>Medicaid enrolment, medical, pharmacy, and care management administrative data from January 1, 2016, to December 31, 2016</td>
<td>Random forest, Support vector machine, Neural network, Decision tree, logistic regression</td>
<td>ROC</td>
<td>Predicting OUD</td>
</tr>
<tr>
<td>(179)</td>
<td>26,982 patients with OUD diagnosis and appropriate Medicaid coverage between 2015 and 2018</td>
<td>Logistic regression, decision tree, random forest, gradient boosting (XGBoost)</td>
<td>Accuracy, AUC, sensitivity, specificity</td>
<td>Overdose events</td>
</tr>
<tr>
<td>(180)</td>
<td>All opioid overdose admitted patients from The Friedman &amp; The Medical College Wisconsin</td>
<td>Logistic regression, random forest, support vector machine classifier, easy ensembles, XGBoost</td>
<td>AUC, accuracy, precision, recall</td>
<td>Length of stay in opioid overdose</td>
</tr>
<tr>
<td>(181)</td>
<td>Substance use population, Age, Sex, Race, Disability, Medicaid, Medicare United States Substance use population</td>
<td>Supervised Regression: Multiple Logistic Regression</td>
<td>AUC</td>
<td>Validating screening risk index for overdose</td>
</tr>
<tr>
<td>(182)</td>
<td>Swedish Youth population, age Swedish Youth population</td>
<td>Supervised Classification: Random Forest, Deep learning, Recurrent neural network</td>
<td>AUROC</td>
<td>Predicting committing suicide or use disorders in ADHD in Youth</td>
</tr>
<tr>
<td>(183)</td>
<td>7915 non-institutionalized US citizens who received at least one opioid prescription between 2016 to 2019</td>
<td>Support vector machine, random forest, neural network, gradient boosting and XGBoost (extreme gradient boosting)</td>
<td>AUROC, AUPRC</td>
<td>Opioid use frequency</td>
</tr>
</tbody>
</table>

Figure 9 Comparative Distribution of Model Types
This systematic review included 69 studies. Of these, 50 used logistic regression, 39 random forest, 27 deep learning, and 21 boosting models, detailed in Figure 9. For evaluation metrics, 56 studies utilized AUC/AUROC, 26 focused on specificity, 25 on sensitivity, and 24 on accuracy, detailed in Figure 10.
### 2.3.1 Meta-Analysis Results

**By Outcome**

<table>
<thead>
<tr>
<th>Author(s) and Year</th>
<th>SMD [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Opioid use disorder (k = 36)</strong></td>
<td></td>
</tr>
<tr>
<td>Wadkar et al., 2020</td>
<td>1.72 [1.54, 1.90]</td>
</tr>
<tr>
<td>Woo-Young Ahna et al., 2016</td>
<td>1.59 [1.37, 1.82]</td>
</tr>
<tr>
<td>Alemi et al., 2018</td>
<td>1.41 [1.27, 1.54]</td>
</tr>
<tr>
<td>Annis et al., 2021</td>
<td>0.46 [0.31, 0.61]</td>
</tr>
<tr>
<td>Banks et al., 2023</td>
<td>1.09 [0.97, 1.21]</td>
</tr>
<tr>
<td>Burgess-Hull et al., 2023</td>
<td>1.59 [1.36, 1.82]</td>
</tr>
<tr>
<td>Charlifsh et al., 2019</td>
<td>2.54 [1.92, 3.16]</td>
</tr>
<tr>
<td>Ciesielski et al., 2016</td>
<td>1.45 [1.29, 1.60]</td>
</tr>
<tr>
<td>Dong et al., 2021b</td>
<td>1.23 [1.12, 1.35]</td>
</tr>
<tr>
<td>Kashyap et al., 2023</td>
<td>1.06 [0.57, 1.55]</td>
</tr>
<tr>
<td>Lo-Ciganic et al., 2020</td>
<td>1.66 [1.41, 1.91]</td>
</tr>
<tr>
<td>Wagner et al., 2021</td>
<td>1.00 [0.77, 1.23]</td>
</tr>
<tr>
<td>Subtotal (Q = 899.47, df = 35, p &lt; 0.001; ( \chi^2 = 0.00 ))</td>
<td>1.42 [1.13, 1.71]</td>
</tr>
</tbody>
</table>

| **Opioid overdose (k = 39)** | |
| Aijmobi et al., 2022 | 0.80 [0.63, 1.14] |
| Badger et al., 2019 | 0.98 [0.76, 1.24] |
| Chang et al., 2020 | 0.97 [0.74, 1.24] |
| Dong et al., 2020 | 1.01 [0.77, 1.24] |
| Ellis et al., 2019 | 1.27 [0.98, 1.56] |
| Ferris et al., 2019 | 1.13 [0.69, 1.57] |
| Lo-Ciganic et al., 2021 | 1.70 [1.37, 2.02] |
| Lo-Ciganic et al., 2022 | 1.36 [1.12, 1.60] |
| Lo-Ciganic et al., 2019 | 1.44 [0.87, 2.00] |
| Saloner et al., 2020 | 1.47 [1.32, 1.63] |
| Sun et al., 2020 | 1.46 [1.08, 1.85] |
| Zedler et al., 2018 | 1.47 [1.02, 1.92] |
| Subtotal (Q = 152.87, df = 38, p < 0.001; \( \chi^2 = 0.00 \)) | 1.37 [1.19, 1.55] |

| **Risk of use (k = 25)** | |
| Aminavadizadeh et al., 2017 | 0.77 [0.34, 1.20] |
| Dong et al., 2021b | 1.11 [1.00, 1.21] |
| Dong et al., 2023 | 1.11 [1.00, 1.21] |
| Unkkil et al., 2021 | 1.37 [0.97, 1.77] |
| Subtotal (Q = 310.73, df = 24, p < 0.001; \( \chi^2 = 0.00 \)) | 1.25 [1.02, 1.49] |

Total (Q = 311749.68, df = 77, p < 0.001; \( \chi^2 = 0.00 \)) 1.28 [1.14, 1.42]

Figure 11 Meta-analysis of machine learning model AUCs in cohort studies, categorized by predictive outcomes
By Model

<table>
<thead>
<tr>
<th>Model Type</th>
<th>k</th>
<th>Subtotal (Q)</th>
<th>df</th>
<th>p</th>
<th>$\tau^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regressions</td>
<td>48</td>
<td>433.87</td>
<td>47</td>
<td>&lt;0.001</td>
<td>0.00</td>
</tr>
<tr>
<td>Decision Trees/Random Forests</td>
<td>32</td>
<td>310.15</td>
<td>31</td>
<td>&lt;0.001</td>
<td>0.00</td>
</tr>
<tr>
<td>Deep Learning models</td>
<td>37</td>
<td>1406.74</td>
<td>36</td>
<td>&lt;0.001</td>
<td>0.00</td>
</tr>
<tr>
<td>Boosting Algorithms</td>
<td>21</td>
<td>289.19</td>
<td>20</td>
<td>&lt;0.001</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**Figure 12 Meta-Analysis of machine learning model AUCs in cohort studies, categorized by model type**
Figure 1 presents effect sizes as standardized mean differences from the AUC of cohort studies, divided by outcomes like opioid use disorder (OUD), opioid overdose (OD), and risk of opioid use. The average effect size across these was 1.28 (95% CI [1.14, 1.42]), from 28 studies covering 100 models. This result showed high heterogeneity but no publication bias per Egger's regression. In subgroup analysis, models for OUD (effect size: 1.42, 95% CI [1.13, 1.71]) and OD (effect size: 1.37, 95% CI [1.19, 1.55]) outperformed those for opioid use risk (effect size: 1.28, 95% CI [1.14, 1.42]).

Figure 12 shows effect sizes as standardized mean differences from the AUC of cohort studies, sorted by model type. The average effect size was 1.27 (95% CI [1.12, 1.42]) from 28 studies with 138 models. Models included logistic regression, decision trees/random forests, deep learning, and boosting algorithms. This estimate was also highly heterogeneous but not prone to publication bias, as per Egger's regression analysis. In terms of subgroup performance, deep learning models had the highest effect size at 1.54 (95% CI [1.15, 1.94]), followed by logistic regression models at 1.22 (95% CI [1.10, 1.34]), decision trees/random forest models at 1.26 (95% CI [1.07, 1.44]), and boosting algorithms at 1.22 (95% CI [1.14, 1.42]).

Lastly, additional forest plots comparing all models and outcomes, and their performance based on AUC can be seen in Appendix A.

2.4 Discussion

As AI has continued to grow, its application in the field of addictions has become increasingly important in assisting with prediction.
2.4.1 Popular models in predicting opioid related outcomes

In this systematic review, all studies employed supervised Machine Learning (ML) models for classifying opioid-related outcomes. They primarily focused on binary or multiclass outcomes using various classification models. This emphasis on supervised learning, particularly suitable for labeled healthcare data, contrasts with unsupervised learning that explores variable relationships without defined outcomes. The studies demonstrated a diverse range of supervised learning models: Logistic Regression appeared in 50 instances, Random Forest/Decision Trees in 39, Deep Learning in 27, and Boosting algorithms in 21. This variety highlights the complex and specific research needs in predicting opioid-related outcomes, with each model offering distinct strengths based on data characteristics and intended outcomes.

Regression models, crucial in many predictive scenarios, were notably underutilized in the reviewed studies. Their absence is intriguing, given their potential for predicting quantitative aspects like opioid consumption amount and treatment duration, which could aid in healthcare cost management and overdose risk forecasting. Regression models, adept at producing continuous values, are ideal for quantifying outcomes. For example, they could estimate opioid dosages for pain management or predict therapy duration for chronic pain, considering patient-specific factors. Integrating these models could significantly enhance personalized medicine, optimizing treatment and reducing opioid-related risks.

2.4.2 Popular metrics in predicting opioid related outcomes

Different studies have adopted various metrics to gauge the performance of their ML models. A recent literature review underlines the complexities in selecting the most apt metrics to evaluate ML model quality, highlighting that popular metrics like accuracy might not be
effective, especially in class-imbalanced datasets (184). Model type (regression vs. classification) dictates the applicable metrics. For classification models, the Area Under the Curve (AUC) emerged as the most common metric (n=56) (111), valued for its comprehensive assessment encompassing a range of probabilities and true/false negatives and positives. Its wide-ranging measurements make AUC a preferred choice for evaluating ML models (111). Specificity, sensitivity, and accuracy (n=26, n=25, n=24 respectively) were also frequently used. Precision and recall gain importance in medical contexts, where patient safety risks like suicide, death, or overdose are concerns. Typically, clinical tools overestimating adverse effects are less harmful than those underpredicting them.

2.4.3 Data-Driven Insights

The development of ML models in addiction psychiatry necessitates careful selection of data sources to reduce biases and enhance precision. This review showcases models incorporating both passively and actively gathered data. Passively sourced data, typically from healthcare services and electronic medical records, offers expansive sample sizes but may lack detailed behavioral information and can be prone to incomplete data. Conversely, actively collected data, obtained through methods like standardized or innovative questionnaires, provides in-depth behavioral insights. Though more resource-intensive, this method ensures the collection of specific data relevant to research questions, enhancing the model's generalizability and effectiveness in addiction psychiatry.

While actively gathered data can achieve a balance, such equilibrium doesn't always accurately reflect specific real-world challenges. For example, skewed class data in datasets might more accurately represent issues prevalent in certain populations with higher fatality rates. Using standardized data like International Classification of Disease (ICD) codes can streamline
preprocessing and enhance model robustness (185). These codes, combined with medical health records such as insurance claims and diagnostic codes, are instrumental in achieving more accurate predictions and robust results in empirical studies (186,187).

2.4.4 Model performance comparing model types: Meta-Analysis Insights

In our comprehensive meta-analysis focusing on models predicting opioid-related outcomes, deep learning models consistently showed the highest effect size, closely followed by boosting algorithms, decision trees/random forests, and logistic regression models. The complexity of overdose outcomes, often entangled with various comorbidities, presents a challenging environment. Deep learning models, due to their capacity to uncover and interpret intricate, non-linear patterns in large and diverse datasets, excel in this context (188). Their sophisticated neural network architectures allow them to effectively learn from and adapt to the complex nature of the data, making them particularly advantageous for analyzing scenarios where conventional linear models are inadequate.

Boosting algorithms and decision trees/random forests also significantly contribute to deciphering complex dynamics (189). Although not as proficient as deep learning in handling highly intricate patterns, they have unique strengths like versatile data handling and model interpretability, essential in clinical contexts. Decision trees provide an intuitive grasp of variable interactions affecting outcomes. Meanwhile, boosting algorithms, through iterative refinement, improve model accuracy in scenarios with unbalanced or noisy data.

Conversely, logistic regression, being a simpler, more linear model, could encounter its limitations in complex scenarios like opioid-related outcome prediction. While it is beneficial for
straightforward tasks and offers ease in interpretation and implementation, its ability to discern the intricate relationships crucial for predicting opioid outcomes is somewhat restricted (190).

While deep learning's effectiveness, as shown by its significant effect size in our meta-analysis, is noteworthy, it's crucial to remember that model selection must be contextual and data driven. The ideal model choice depends on factors like data availability, quality, and the specific requirements of the research question. Our review often observed custom models tailored to specific research challenges, emphasizing the need for bespoke approaches to address the complexities of opioid-related outcomes. These custom models are crafted to suit the unique data aspects and specific research questions, highlighting the value of a tailored approach. This analysis showcases the variety of tools available for addiction psychiatry research and underscores the need for a nuanced approach in model selection. It emphasizes balancing theoretical effectiveness with practicality and the specific demands of the data. This comprehensive perspective is vital in advancing the field's research methodologies.

2.4.5 Model performance comparing model outcomes: Meta-Analysis Insights

Our review pinpointed four main predictive outcomes in opioid-related research: Opioid Use Disorder (OUD), Opioid Overdose, Opioid Use, and Risk of Opioid Use. These outcomes are essential in comprehending and tackling various aspects of the opioid epidemic. Models that focus on reducing opioid use risk or influencing treatment paths can significantly impact behavioral changes, potentially shifting the trajectory of the crisis.

Our analysis revealed that studies focusing on opioid overdose generally produced more consistent results compared to those examining Opioid Use Disorder (OUD), which showed greater variability. This suggests that OUD is influenced by a more complex array of variables,
underscoring the need for further research to better understand these factors. The heterogeneity observed in OUD studies implies potential overlooked covariates, which could offer deeper insights into the disorder's nature.

The intricacy of opioid-related issues highlights the necessity of precisely defining predictive outcomes in research. The choice between focusing on opioid use likelihood, overdose risk, or Opioid Use Disorder (OUD) intricacies is crucial for developing effective models. These decisions fundamentally shape the models, which are vital for addressing practical challenges in addiction psychiatry.

Our review showed that models predicting Opioid Use Disorder and overdose had higher standardized mean differences than those predicting usage risk. This suggests that the current understanding of 'risk of use' might be unclear or incomplete, affecting data collection and prediction accuracy. This lack of clarity underscores the necessity for more detailed research into the risk factors of opioid use, crucial for improving model precision in addiction psychiatry and aiding in more effective interventions and policy development.

2.4.6 Future of Machine learning: Harnessing big data/AI and its limitations

Machine Learning (ML) technology has the extraordinary ability to process large data volumes, providing insights that could revolutionize addiction treatment and research. Traditional methods in addiction medicine often depend on fixed structures and protocols, which can limit flexibility and innovation. ML, however, offers a dynamic, evidence-based approach, potentially uncovering new solutions across the spectrum of addiction treatment. This represents a significant shift from conventional methodologies, enabling more responsive and adaptive strategies in the field.
The flexibility of ML technology enables its use across various stages of addiction treatment, including prevention, harm reduction, treatment, and maintenance. This adaptability is exemplified in predictive modeling for risk behavior (191), forecasting drug treatment effectiveness (192), and predicting adherence to treatment protocols (193). These applications demonstrate ML's potential to significantly impact the course of addiction treatment.

Utilizing Machine Learning (ML) in psychiatry has notable benefits, but it's important to consider its limitations. A primary concern is data quality and inherent biases, which can be amplified by ML's iterative process. The challenge lies not just in applying ML to datasets, but in interpreting and applying these models in real-world contexts. The success of ML models largely depends on the data they're based on. The quality and scope of the data may not always encapsulate all relevant real-world variables, affecting the model's outcomes.

A notable limitation of ML models is their tendency to minimize mean squared error, which can inadvertently perpetuate biases (194,195). These biases fall into three categories: sample bias, prejudice bias, and group attribution bias. Sample and group attribution biases may arise from datasets that inaccurately generalize individual traits to a whole group, leading to models that miss deviations. Prejudice bias often originates from researchers' subjective biases, affecting data used in models. Additionally, an algorithm's generalizability is limited to its training data, possibly excluding essential risk factors that are not presented in the dataset.

To address these limitations, it's vital to have a comprehensive methodology for evaluating ML models' real-world utility and efficacy. Increasing model flexibility and application can be achieved through universal coding or standardized questionnaires. These strategies help encompass a wider range of variables and minimize training and application biases in ML models.
2.5 Conclusion

Machine Learning (ML) in addiction psychiatry leverages big data to identify addiction-related factors, aiding in creating personalized treatment approaches. Research primarily uses classification models, focusing on substance use, mental health, demographics, and clinical markers. Evaluation metrics like the area under the receiver operating characteristic curve, accuracy, sensitivity, and specificity are common, particularly for overdose outcomes. These insights highlight ML's vital role in enhancing addiction psychiatry. However, further research is needed to improve data collection and model development, especially for underrepresented groups and including critical variables. Ethical considerations in practical healthcare applications of these models are also crucial.

2.5.1 Study Limitations

Our systematic review faces several limitations. Focusing solely on English-language studies from certain databases might lead to selection bias. The variety in study populations, risk factors, ML models, and metrics across studies complicates quantitative comparisons. This diversity, along with occasional gaps in reporting model evaluations and feature selection, limits our ability to draw consistent, direct comparisons across the studies.
## 2.6 Risk of Bias

Table 2 Risk of bias assessment for COHORT studies (196) for Machine Learning in Drug toxicity crisis Management: A Systematic Review and Meta-Analysis

<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Modified Newcastle Ottawa Quality Assessment Scale for COHORT studies</th>
<th>Score (/9)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>REPRESENTATIVENESS OF THE EXPOSED COHORT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>At least somewhat representative of the exposed cohort</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Drawn from same community as exposed cohort</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Structured interview or secure record</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Afshar et al., 2021 (118)</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Ahn &amp; Vassileva, 2016 (119)</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Ajumobi et al., 2022 (120)</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Alemi et al., 2018 (121)</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Amirabadi Zadeh et al., 2018 (122)</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Annis et al., 2022 (123)</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Badger et al., 2019 (125)</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>SELECTION (max 4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Selection of the non-exposed cohort</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>COMPARABILITY (max 2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adjusts for most important risk factor</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OUTCOME (max 3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Assessment of outcome</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Was follow-up long enough for outcomes to occur?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adequacy of follow-up of cohorts</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Complete follow-up or subjects lost unlikely to introduc e bias</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Score (/9)</td>
<td></td>
</tr>
<tr>
<td>Afshar et al., 2021 (118)</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Ahn &amp; Vassileva, 2016 (119)</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Ajumobi et al., 2022 (120)</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Alemi et al., 2018 (121)</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Amirabadi Zadeh et al., 2018 (122)</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Annis et al., 2022 (123)</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Badger et al., 2019 (125)</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Year(s)</td>
<td>Value(s)</td>
</tr>
<tr>
<td>-----------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>Banks et al., 2023</td>
<td>2023</td>
<td>(126)</td>
</tr>
<tr>
<td>Beattie &amp; Nicholson, 2021</td>
<td>2021</td>
<td>(127)</td>
</tr>
<tr>
<td>Burgess-Hull et al., 2022</td>
<td>2022</td>
<td>(128)</td>
</tr>
<tr>
<td>Burgess-Hull et al., 2023</td>
<td>2023</td>
<td>(129)</td>
</tr>
<tr>
<td>Calcaterra et al., 2018</td>
<td>2018</td>
<td>(130)</td>
</tr>
<tr>
<td>Calcaterra et al., 2023</td>
<td>2023</td>
<td>(131)</td>
</tr>
<tr>
<td>Chang et al., 2020</td>
<td>2020</td>
<td>(132)</td>
</tr>
<tr>
<td>Chartash et al., 2019</td>
<td>2019</td>
<td>(133)</td>
</tr>
<tr>
<td>Che et al., 2018</td>
<td>2018</td>
<td>(134)</td>
</tr>
<tr>
<td>Choi &amp; Boo, 2020</td>
<td>2020</td>
<td>(135)</td>
</tr>
<tr>
<td>Ciesielski et al., 2016</td>
<td>2016</td>
<td>(136)</td>
</tr>
<tr>
<td>Cox et al., 2020</td>
<td>2020</td>
<td>(137)</td>
</tr>
<tr>
<td>Dong et al., 2020</td>
<td>2020</td>
<td>(138)</td>
</tr>
<tr>
<td>Dong et al., 2023</td>
<td>2023</td>
<td>(139)</td>
</tr>
<tr>
<td>Dong, Deng, Hou, et al., 2021</td>
<td>2021</td>
<td>(140)</td>
</tr>
<tr>
<td>Dong, Deng, Rashidian, et al., 2021</td>
<td>2021</td>
<td>(141)</td>
</tr>
<tr>
<td>Ellis et al., 2019</td>
<td>2019</td>
<td>(142)</td>
</tr>
<tr>
<td>Ferris et al., 2019</td>
<td>2019</td>
<td>(143)</td>
</tr>
<tr>
<td>Ferris et al., 2021</td>
<td>2021</td>
<td>(144)</td>
</tr>
<tr>
<td>Fojo et al., 2021</td>
<td>2021</td>
<td>(145)</td>
</tr>
<tr>
<td>Fulton et al., 2019</td>
<td>2019</td>
<td>(146)</td>
</tr>
<tr>
<td>Gellad et al., 2023</td>
<td>2023</td>
<td>(147)</td>
</tr>
<tr>
<td>Glanz et al., 2018</td>
<td>2018</td>
<td>(148)</td>
</tr>
<tr>
<td>Gong et al., 2021</td>
<td>2021</td>
<td>(149)</td>
</tr>
<tr>
<td>Gottlieb et al., 2022</td>
<td>2022</td>
<td>(150)</td>
</tr>
<tr>
<td>Guo et al., 2021</td>
<td>2021</td>
<td>(151)</td>
</tr>
<tr>
<td>Han et al., 2020</td>
<td>2020</td>
<td>(152)</td>
</tr>
<tr>
<td>Hasan et al., 2021</td>
<td>2021</td>
<td>(153)</td>
</tr>
<tr>
<td>Hu et al., 2020</td>
<td>2020</td>
<td>(154)</td>
</tr>
<tr>
<td>Reference</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>Hylan et al., 2015 (155)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Kashyap et al., 2023 (156)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Liu et al., 2023 (157)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Lo-Ciganic et al., 2019 (158)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Lo-Ciganic et al., 2020 (159)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Lo-Ciganic et al., 2021 (160)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Lo-Ciganic et al., 2022 (161)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>McDougall et al., 2020 (163)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Nath et al., 2017 (164)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Oliva et al., 2022 (165)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Prince et al., 2020 (166)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Rakovski et al., 2023 (168)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Rekabdar et al., 2022 (167)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Rice et al., 2012 (169)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Saloner et al., 2020 (170)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Schell et al., 2022 (171)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Segal et al., 2020 (172)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Sharma et al., 2020 (124)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Sun et al., 2020 (173)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Tabar et al., 2020 (174)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Tong et al., 2021 (175)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Tseregounis et al., 2021 (176)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Vunikili et al., 2021 (177)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>A. Wadekar, 2019 (116)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>A. S. Wadekar, 2020 (117)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Wagner et al., 2021 (178)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Wanzhen Gao et al., 2021 (179)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Warren et al., 2022 (180)</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

62
<table>
<thead>
<tr>
<th>Study</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wu et al., 2021 (181)</td>
<td>0</td>
</tr>
<tr>
<td>Zedler et al., 2018 (182)</td>
<td>1</td>
</tr>
<tr>
<td>Zong et al., 2020 (61)</td>
<td>0</td>
</tr>
<tr>
<td>Zhao et al., 2021 (183)</td>
<td>1</td>
</tr>
</tbody>
</table>
Chapter 3: Clinical decision support systems in addictions and concurrent disorders: a systematic review and meta analysis

3.1 Introduction

Addictions and concurrent disorders significantly contribute to the global disease burden. Their impact extends beyond affected individuals, affecting families, relationships, and communities. These disorders present complex management challenges due to their socio-economic, psychosocial, and cultural effects. The impact of substance use disorders since March 2020 has been comparable to COVID-19's effects, with 13.83 million Disability-Adjusted Life-Years (DALYs) attributed to substance use versus 15.03 million DALYs for COVID-19 (197).

The global recognition of addictions and concurrent disorders highlights the need for holistic treatment approaches, tailored to individual needs. Effective strategies should be informed by large-scale data and behavioral patterns (198). A comprehensive, multidimensional approach, involving healthcare systems, policymakers, and society, is essential (199). This strategy should focus on evidence-based solutions, supporting decision-making, increasing safety and prevention, and ultimately improving the quality of life for those affected by these conditions (200).

The complexity of addiction and concurrent disorders presents substantial challenges in clinical decision-making due to each patient's unique combination of symptoms, history, and comorbidities, which demand individualized therapeutic approaches. However, the clinical field is currently facing significant gaps, notably a marked shortage of healthcare professionals (201). This issue is exacerbated by the growing number of patients, overdoses, and crises, leading to overstretched healthcare infrastructure. Overburdened healthcare professionals often struggle with a vast array of guidelines and recommendations. Despite the availability of extensive
information, the application of these guidelines is often inconsistent, leading to variable clinical judgments and sometimes resulting in unnecessary or even potentially detrimental diagnostic tests and treatments (202).

3.1.1 What is a Clinical Decision Support System?

A Clinical Decision Support System (CDSS) is a technology developed to enhance the accuracy and effectiveness of clinical decisions. It integrates clinical expertise, patient history, and ongoing medical research, providing healthcare professionals with actionable and informed insights. This software is specifically designed to support clinicians and health professionals in making well-informed decisions at the point of care, thereby improving patient outcomes and optimizing healthcare processes (203,204).

The CDSS process starts by collecting specific patient data, like symptoms, medical history, or diagnostic results. This data is compared to a comprehensive, computerized clinical knowledge base, continually updated with the latest medical research and expert guidelines (204,205). The system then aligns the patient's information with this vast database, generating personalized assessments or recommendations tailored to each patient's unique medical background (206).

Contemporary CDSS have progressed to process intricate data sets, thereby significantly enhancing personalized care. These systems offer various benefits, from equipping healthcare professionals with critical insights to tailoring patient-centric clinical pathways that adapt treatments to specific needs. In dynamic clinical environments, CDSS serve as crucial connectors, ensuring patients consistently receive key medical services (204). CDSS are typically divided into knowledge-based systems, which analyze patient data using established
rules and inference mechanisms, and non-knowledge-based systems that utilize artificial intelligence and machine learning to evolve and identify patterns in extensive clinical data (204). Their capacity to adapt and stay current with the latest treatments and best practices positions them as essential tools in addressing challenges like healthcare workforce shortages (207,208).

The importance of Clinical Decision Support Systems (CDSS) is especially highlighted in addiction treatment. These systems harness data from individual user platforms to forecast psychiatric developments, including the likelihood of relapses or overdoses, thereby functioning as tools for both prediction and prevention (209,210). Additionally, they are crucial in preventing the overlook of information due to time constraints in clinical practice (211), in reducing biases, and in enhancing diagnosis in settings with limited resources (209). Furthermore, the integration of Electronic Medical Records (EMRs) with non-knowledge-based CDSS enables the effective analysis and interpretation of a range of data, from vital signs to behavioral patterns, ensuring timely psychiatric intervention ahead of a crisis (212,213). Advancing CDSS into current online infrastructures represents a significant development for their continuous evolution and integration. These platforms enable real-time data collection, focus on patient-centered care, and improve the efficiency of specific referrals. This forward-thinking transformation in the digital health sector not only ensures timely and relevant care but also promotes a more integrated healthcare ecosystem. Prominent examples of this development are the Innowell platform from the University of Sydney and the Risk Assessment and Management Platform from the University of British Columbia (214,215).

Although there's significant potential, detailed studies on the use of Clinical Decision Support Systems (CDSS) specifically for addiction and concurrent disorders are rare. This systematic review aims to fill this gap by examining how CDSS affects the treatment and
outcomes of these disorders. The goal is to fully comprehend the impact of CDSS, shedding light on future treatment approaches and support systems for individuals grappling with these multifaceted conditions.

3.2 Methods

3.2.1 Protocol

The protocol for this study was developed in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (104), and was further refined through consultations with the Cochrane Handbook for Systematic Reviews and Interventions (105). It was subsequently registered with Prospero (216). Given the nature of the study, seeking approval from an ethics committee was deemed unnecessary, as outlined in the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (107).

3.2.2 Search Strategy

This research involved comprehensive searches of several databases from their start dates up to May 25, 2023. These databases included MEDLINE (Ovid), Embase (Ovid), CINAHL (Ebsco), PsycINFO (Ebsco), and Web of Science, accessed through the University of British Columbia. Subject experts (AT, JK) collaborated closely with a medical research librarian (VK) to devise a detailed list of subject headings and keywords relevant to each aspect of the research question. The librarian (VK) then tailored all search strategies based on the approach developed for MEDLINE, using a similar set of terms. These searches were refinements of an initial search
conducted by the subject experts (AT, JK). Details of the search strategies can be found in Appendix B.1.

3.2.3 Eligibility Criteria

Inclusion criteria for the studies focused on those addressing the description, assessment, or implementation of clinical decision support systems (CDSS) in the context of substance abuse and concurrent disorders. The CDSS in these studies needed to offer personalized feedback to either healthcare providers or the individuals concerned. All types of studies were considered, except for conference proceedings, editorials, text, and opinion papers. Exclusion criteria encompassed studies involving FMRI/Imaging, Natural Language Processing (NLP), topics in general medicine and primary care related to pain, shared decision-making processes without the use of a CDSS tool, or studies focused on postpartum depression.

3.2.4 Risk of Bias

The assessment of bias risk in the studies was independently conducted by two reviewers (AT, JW). Reviewers utilized RoB 2: A revised Cochrane risk-of-bias tool for randomized trials, Modified Newcastle Ottawa Assessment Scale for (Cohort Studies, Randomized Controlled Trials, Case Controls, Cross Sectional Studies), and the CASP tool for qualitative appraisal. In cases of disagreement on assessed items, resolution was sought by consulting a third reviewer (JK) for their opinion.

3.2.5 Data Extraction

Two reviewers (AT, JK) independently conducted data extraction and assessed the risk of bias for each included study. The data they extracted covered various aspects, such as study
design, demographics, the type of CDSS used, user outcomes, clinical outcomes, model maturity, and the focus of the study. In cases where specific data points were not clearly mentioned, those studies were excluded from the meta-analysis but included for qualitative data extraction. The data extraction form was designed beforehand.

We aimed for an exhaustive search of relevant studies, but it's important to recognize that language and geographic location could have introduced bias in our research.

3.2.6 **Criteria for classification for qualitative analysis**

To assess the maturity of Clinical Decision Support System (CDSS) models, Simon's iterative decision-making process phases was utilized, a method previously applied in systematic reviews (217). Four stages were defined based on Simon's phases to indicate CDSS maturity (218):

Stage 1: Intelligence - This initial phase is centered on gathering data and identifying problems.

Stage 2: Design + Intelligence - This stage involves the development of algorithms and interfaces, building upon the intelligence gathered.

Stage 3: Choice + Design + Intelligence - At this stage, thorough testing is conducted to select the most effective system.

Stage 4: Implementation + Choice + Design + Intelligence - The final stage, where the system is deployed in real-world clinical settings and subject to ongoing evaluation.

These stages collectively offer a comprehensive view of a CDSS's development and preparedness for clinical use (218). They provide a structured approach to creating data-informed, clinically relevant decision-making tools (218).
3.2.7 Methodology for Meta-analysis

Meta-analyses were carried out on patient outcome measures from randomized control trials (RCTs), encompassing efficacy, satisfaction, and acceptance metrics. Recognizing that each Randomized Controlled Trial (RCT) adapted its estimates to various covariates, it was concluded that unadjusted estimates would offer a more uniform basis for comparison than adjusted ones. Therefore, all raw estimates were included in the meta-analyses. Dependencies arising from this were addressed through aggregation and multi-level modeling techniques. Specifically, if an outcome measure (like depression) was reported at different times (e.g., 1 month, 6 months, 12 months), these were aggregated. Furthermore, when a single study reported multiple relevant outcomes (such as depression and anxiety), three-level models were employed to account for dependencies, incorporating all pertinent outcomes from each study.

Three distinct analyses were conducted. The first encompassed all measures, while the second and third analyses were subgroup analyses focusing solely on efficacy and acceptance measures, respectively. The impact of subgrouping was assessed using meta-regression. All analyses employed Standardized Mean Differences (SMD) and were performed using R, Version 4.3.1 (219), with the metafor package V. 4.2-0 (115).
3.3 Results

Figure 13 CDSS PRISMA flow diagram (104)
Table 3 Qualitative Analysis of studies for for Machine Learning in Drug toxicity crisis Management: A Systematic Review and Meta-Analysis (Papers highlighted with the same color describe different phases of the same study)

<table>
<thead>
<tr>
<th>Study</th>
<th>Study Design</th>
<th>Data Source</th>
<th>Type of CDSS</th>
<th>Results (Clinical)</th>
<th>Study Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Randomized Controlled Trial</td>
<td>Interventions: 17 staff members in psychiatry</td>
<td>Knowledge-based</td>
<td>Decreased number of treatment-related visits and shorter duration of stay</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>2</td>
<td>Randomized Controlled Trial</td>
<td>Control Trial: 12 staff members in psychiatry</td>
<td>Knowledge-based</td>
<td>Increased knowledge about adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>3</td>
<td>Randomized Controlled Trial</td>
<td>Intervention: 17 patients in experimental group</td>
<td>Knowledge-based</td>
<td>Improved adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>4</td>
<td>Randomized Controlled Trial</td>
<td>Control Trial: 12 patients in the same group</td>
<td>Knowledge-based</td>
<td>No significant difference in adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>5</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>6</td>
<td>Randomized Controlled Trial</td>
<td>Control physicians in the study</td>
<td>Knowledge-based</td>
<td>No significant difference in adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>7</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>8</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>9</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>10</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>11</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>12</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>13</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>14</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>15</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>16</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>17</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>18</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>19</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>20</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>21</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>22</td>
<td>Randomized Controlled Trial</td>
<td>Participating physicians in the study</td>
<td>Knowledge-based</td>
<td>Increased adherence to treatment guidelines</td>
<td>Global Survival Rate</td>
</tr>
<tr>
<td>Title</td>
<td>Study Design</td>
<td>Setting</td>
<td>Type of CDS</td>
<td>Measure</td>
<td>Outcome</td>
</tr>
<tr>
<td>-------</td>
<td>--------------</td>
<td>---------</td>
<td>-------------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>Clinical</td>
<td>Pilot study</td>
<td>15 total, 13 physicians and 2 advanced nurse practitioners</td>
<td>Knowledge based</td>
<td>Residents were surveyed for IDDEAS tool and ASHAs</td>
<td>Stage 3</td>
</tr>
<tr>
<td>Intervention</td>
<td>Randomized controlled trial</td>
<td>154 patient days of records from hospital's database</td>
<td>Knowledge based</td>
<td>IDDEAS prototype usability</td>
<td>Stage 3</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>66 individuals with opioid use disorder</td>
<td>Knowledge based</td>
<td>Stepwise approach developed a protocol of a mental health-specific IDDEAS. Clinicians receive a two-component protocol of medication history and medication check.</td>
<td>Stage 3</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>31 total, 0.55 compared with 0.29) as medication management visits</td>
<td>Knowledge based</td>
<td>Computer Automation (CHICA) system is integrated in daily workflow. Further usability assessments and identification of additional IDDEAS requirements are necessary</td>
<td>Stage 2</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>30 child and adolescent health datasets</td>
<td>Knowledge based</td>
<td>IDDEAS will also use Norway’s resources</td>
<td>Stage 2</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>40 adults with chronic suicidal thoughts</td>
<td>Knowledge based</td>
<td>First 20 individuals from the age of 6 to 12, with or without treatment for a first medication history and guidance on medication choice.</td>
<td>Stage 2</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>40% vs UC, 75%; P = .02), it also improved clinicians' decisional comfort (DMC, 80% vs UC, 68%; P &lt; .001) and satisfaction (80% vs UC, 75%; P = .02)</td>
<td>Knowledge based</td>
<td>Offered clinical problem solving in 2 clinics</td>
<td>Stage 1</td>
</tr>
<tr>
<td>Intervention</td>
<td>Clinical Decision Making</td>
<td>776 participants completed to week 8 are required. Recruitment started in the UK in July 2016. Recruitment in other countries of interest were randomized using population-based recruitment.</td>
<td>Knowledge based</td>
<td>Clinical Decision Making</td>
<td>Stage 1</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>15.0 [14.0; 16.0]</td>
<td>Knowledge based</td>
<td>Drug withdrawal state</td>
<td>Stage 1</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>13.5 [11.2; 16.0]</td>
<td>Knowledge based</td>
<td>Drug withdrawal state</td>
<td>Stage 1</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>2.8% (8/288) in the pre-intervention phase</td>
<td>Knowledge based</td>
<td>Guided emergency department clinicians to initiate buprenorphine/naloxone treatment and initiate follow-up recommendations</td>
<td>Stage 2</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>50% (8/16) vs 25% (4/16);</td>
<td>Knowledge based</td>
<td>Computer Automated (CHICA) system</td>
<td>Stage 2</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>0.55 compared with 0.29) as medication management visits</td>
<td>Knowledge based</td>
<td>Computer Automated (CHICA) system</td>
<td>Stage 2</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>80% vs UC, 75%; P = .02), it also improved clinicians' decisional comfort (DMC, 80% vs UC, 68%; P &lt; .001) and satisfaction (80% vs UC, 75%; P = .02)</td>
<td>Knowledge based</td>
<td>Offered clinical problem solving in 2 clinics</td>
<td>Stage 1</td>
</tr>
<tr>
<td>Intervention</td>
<td>Clinical Decision Making</td>
<td>776 participants completed to week 8 are required. Recruitment started in the UK in July 2016. Recruitment in other countries of interest were randomized using population-based recruitment.</td>
<td>Knowledge based</td>
<td>Clinical Decision Making</td>
<td>Stage 1</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>15.0 [14.0; 16.0]</td>
<td>Knowledge based</td>
<td>Drug withdrawal state</td>
<td>Stage 1</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>13.5 [11.2; 16.0]</td>
<td>Knowledge based</td>
<td>Drug withdrawal state</td>
<td>Stage 1</td>
</tr>
<tr>
<td>Intervention</td>
<td>Clinical Decision Making</td>
<td>776 participants completed to week 8 are required. Recruitment started in the UK in July 2016. Recruitment in other countries of interest were randomized using population-based recruitment.</td>
<td>Knowledge based</td>
<td>Clinical Decision Making</td>
<td>Stage 1</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>50% (8/16) vs 25% (4/16);</td>
<td>Knowledge based</td>
<td>Computer Automated (CHICA) system</td>
<td>Stage 2</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>0.55 compared with 0.29) as medication management visits</td>
<td>Knowledge based</td>
<td>Computer Automated (CHICA) system</td>
<td>Stage 2</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>80% vs UC, 75%; P = .02), it also improved clinicians' decisional comfort (DMC, 80% vs UC, 68%; P &lt; .001) and satisfaction (80% vs UC, 75%; P = .02)</td>
<td>Knowledge based</td>
<td>Offered clinical problem solving in 2 clinics</td>
<td>Stage 1</td>
</tr>
<tr>
<td>Intervention</td>
<td>Clinical Decision Making</td>
<td>776 participants completed to week 8 are required. Recruitment started in the UK in July 2016. Recruitment in other countries of interest were randomized using population-based recruitment.</td>
<td>Knowledge based</td>
<td>Clinical Decision Making</td>
<td>Stage 1</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>50% (8/16) vs 25% (4/16);</td>
<td>Knowledge based</td>
<td>Computer Automated (CHICA) system</td>
<td>Stage 2</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>0.55 compared with 0.29) as medication management visits</td>
<td>Knowledge based</td>
<td>Computer Automated (CHICA) system</td>
<td>Stage 2</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>80% vs UC, 75%; P = .02), it also improved clinicians' decisional comfort (DMC, 80% vs UC, 68%; P &lt; .001) and satisfaction (80% vs UC, 75%; P = .02)</td>
<td>Knowledge based</td>
<td>Offered clinical problem solving in 2 clinics</td>
<td>Stage 1</td>
</tr>
<tr>
<td>Intervention</td>
<td>Clinical Decision Making</td>
<td>776 participants completed to week 8 are required. Recruitment started in the UK in July 2016. Recruitment in other countries of interest were randomized using population-based recruitment.</td>
<td>Knowledge based</td>
<td>Clinical Decision Making</td>
<td>Stage 1</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>50% (8/16) vs 25% (4/16);</td>
<td>Knowledge based</td>
<td>Computer Automated (CHICA) system</td>
<td>Stage 2</td>
</tr>
<tr>
<td>Intervention</td>
<td>Intervention</td>
<td>0.55 compared with 0.29) as medication management visits</td>
<td>Knowledge based</td>
<td>Computer Automated (CHICA) system</td>
<td>Stage 2</td>
</tr>
</tbody>
</table>
In this systematic review, 69 studies were analyzed. Among them, 55 employed knowledge-based systems, 11 utilized non-knowledge-based systems, and 4 incorporated both types. Until 2017, the focus was exclusively on knowledge-based models, but from 2017 to 2022, non-knowledge models began to emerge (as illustrated in Figure 14). Of these studies, 29 specifically addressed individual mental health conditions, including psychosis, suicide, PTSD, depression, and ADHD. Other areas of focus were general mental health (21 studies), substance use (16 studies), and adolescent mental health (4 studies). The functions of the CDSS in these studies ranged from aiding in diagnosis and monitoring to managing medication and follow-up care.

In terms of model maturity, 2 models were classified as Stage 1, 27 as Stage 2, 8 as Stage 3, and 32 as Stage 4. Detailed information about the included studies is presented in Table 2.
Figure 14 Comparative Analysis of Publication Count and Types of Clinical Decision Support Systems by Year
3.3.1 Meta-Analysis Results

**Figure 15 Meta-analysis of efficacy/satisfaction & acceptance of CDSS in RCTs included in this systematic review.**

Figure 15 illustrates the effect sizes in Randomized Controlled Trials (RCTs) using CDSS, measured as the standardized mean difference between control and intervention groups.

For efficacy outcomes, the average effect size was -0.11 (95% Confidence Interval [CI]: [-0.21, -0.01]), derived from 8 studies covering 21 outcomes. This result exhibited high heterogeneity ($I^2 = 87.2\%$) but was not influenced by publication bias as indicated by Egger's regression analysis.
In contrast, patient satisfaction and acceptance linked to CDSS usage showed an effect size of -0.50 (95% CI: [-0.33, -0.08]). This outcome, also heterogeneous ($I^2 = 78.2\%$), was subject to publication bias, according to Egger's regression analysis. A meta-regression indicated a significant difference between outcomes related to efficacy and those associated with satisfaction and acceptance ($p < .01$).

### 3.4 Discussion

The continuous evolution of technology in healthcare is becoming more critical, especially in areas that require complex decision-making. The global shortage of physicians, limited funding, and increasing stress in healthcare settings add to the complexity of clinical decision-making. Clinical Decision Support Systems (CDSS) are increasingly vital, offering substantial support to patients and aiding healthcare providers in challenging tasks across the entire spectrum of care (289). These systems integrate diverse clinical data, providing healthcare professionals with comprehensive and relevant information to make informed decisions. CDSS have demonstrated their effectiveness by enhancing decision-making processes, promoting adherence to guidelines, minimizing human errors, and improving patient outcomes (290–297).

The complex nature of addiction and concurrent disorders makes them prime candidates for benefiting from technological advancements. In the field of psychiatry, where challenges such as the lack of objective diagnostic tools, resource constraints, and physician shortages are prevalent, Clinical Decision Support Systems (CDSS) become an essential tool (298). CDSS aids clinicians by providing vital information, enhancing the accuracy of diagnoses and treatment decisions through both knowledge-based and non-knowledge-based approaches.
The essence of any clinical decision support system (CDSS) lies in its knowledge representation (299). Knowledge-based systems, which are grounded in structured data and specific rules, provide clear and actionable insights and recommendations personalized to each patient's profile (299). To stay current and align with evolving treatment protocols and expertise, these systems need regular updates. In contrast, non-knowledge-based systems use algorithms and neural networks to analyze extensive datasets (299). While these systems can deliver profound insights, they sometimes may not offer the clarity and transparency characteristic of knowledge-based systems.

3.4.1 From Knowledge-Based to Data-Driven: CDSS Evolution

The qualitative analysis reveals a significant trend in the realm of Clinical Decision Support System (CDSS) publications over three decades. Historically, these studies predominantly concentrated on knowledge-based systems, which were based on expert opinions and structured assessments. Yet, as depicted in Figure 14, there was a noticeable pivot in 2017 towards embracing non-knowledge-based systems, emphasizing data-driven approaches.

This shift over time reflects a larger trend in the accumulation of data and the rise of big data, accompanied by a concurrent growth in computational power facilitating artificial intelligence and machine learning techniques. This period signifies a pivotal change in research methodologies within the field of addictions, evidenced by the prevalent use of data-driven systems after 2017 (300).

The current trends in data collection and the rapid progress in computational technology suggest that the use of non-knowledge-based systems in Clinical Decision Support Systems (CDSS), especially in addiction and concurrent disorders, is likely to keep growing. This
trajectory is in line with the emerging focus on precision medicine and tailored healthcare interventions, marking a new era in healthcare decision-making (301). However, it's important to recognize that the integration and practical use of these advanced CDSSs might be unfolding more slowly than expected. This slower pace could be attributed to the challenges in altering clinical practices, alongside reservations, mistrust, and uncertainties among healthcare practitioners. Identifying and understanding these barriers could provide crucial insights into the factors limiting the wider adoption of these systems.

3.4.2 Assessing the Maturity of Clinical Decision Support System Models

The findings from our review reveal that most of the Clinical Decision Support System (CDSS) models examined were predominantly in either Stage 2 (Design) or Stage 4 (Implementation). This reflects the current state of maturity in CDSS development. The concentration of models in these stages indicates an ongoing evolution within the field. Stage 2 focuses on the development of theoretical frameworks, algorithms, and interfaces, whereas systems in Stage 4 are either in the process of real-world clinical testing or are already being applied in clinical environments (218). This split between stages points to a scenario where many models are in the initial stages of ideation and development, while others have advanced to more mature phases of implementation.

The results of this study suggest that Simon's framework can act as an essential guide for both researchers and clinicians involved in the development or deployment of Clinical Decision Support Systems (CDSS) (217). For systems currently in the Design stage, the subsequent step is to progress to the Choice stage. This involves assessing their efficacy via thorough clinical trials or comparative studies. For systems in the Implementation stage, the focus should be on
continuous monitoring, gathering feedback, and periodic updates to address real-world challenges and maintain optimal functionality. Simon's framework thus provides a dual benefit: a means for evaluation and a structured approach for the advancement and refinement of CDSS (217). As the field progresses, the relevance and utility of this framework in ensuring the effectiveness and clinical applicability of these systems are expected to grow.

3.4.3 CDSS Efficacy and Addiction Outcomes: Meta-Analysis Insights

The outcomes of the meta-analysis, as shown in Figure 15, indicate that although the efficacy of Clinical Decision Support Systems (CDSS) is statistically significant, its impact may not be substantial enough to markedly alter addiction treatment outcomes. This is suggested by an effect size of -0.11, which implies a relatively modest effect. Nevertheless, the slight immediate impact of CDSS on addiction could be offset by the increased satisfaction and acceptance observed among patients, which is indicated by a more notable effect size of -0.5. This could potentially lead to more significant changes in the long term.

In medical decision-making, the trust that healthcare professionals have in their patients and the tools used for clinical evaluations is crucial. Our research suggests that while the current effect of Clinical Decision Support Systems (CDSS) on patient outcomes is moderate, this is expected to improve as clinicians become more accustomed and skilled in using these sophisticated systems. As healthcare providers deepen their understanding of the algorithms and decision processes within CDSS, their reliance on and confidence in these tools are likely to increase. Presently, the individual-level impact might be modest, partly due to the systems' limited personalization capabilities. The expectation is that as treatment methods become
increasingly tailored to individual needs, there will likely be an improvement in patient outcomes.

The anticipated rise in practitioner confidence in Clinical Decision Support Systems (CDSS) is likely to lead to their more frequent integration into clinical workflows, thereby enhancing their effectiveness in patient care. Supporting this, our study highlights high levels of patient satisfaction and acceptance in interactions with CDSS. This cycle of growing confidence and increased usage likely explains the statistical effect sizes observed in our research, pointing to a positive future for the broader and more efficient use of CDSS in healthcare.

However, our findings also underscore the need for further research. Several factors, including clinicians' judgment and potential biases, can influence how CDSS are integrated, their success, and acceptance in clinical practice. These elements need comprehensive assessment before CDSS can be widely implemented. It's crucial, particularly for clinicians hesitant to change established procedures or adopt new technologies, to recognize the importance of this evolving landscape in healthcare. Although the immediate and transformative effects of Clinical Decision Support Systems (CDSS) on addiction outcomes might not be readily apparent, their capacity to significantly alter the wider clinical environment is clear. This highlights the importance of continuous research into medical decision-making, the progressive functions of CDSS, and the various elements that affect their incorporation into healthcare practices.

3.4.4 Advancements and Future of CDSS: Enhancing Decision-Making and Adaptability

As technology continues to advance, Clinical Decision Support Systems (CDSS) are expected to grow in their effectiveness and efficiency. A critical aspect of this progression
involves ensuring that these systems are inclusive and accessible to marginalized populations. This approach is vital for enhancing the overall accessibility of care and addressing disparities in healthcare access (302).

A robust Clinical Decision Support System (CDSS) is essential for accurately processing and interpreting symptoms, which is crucial in providing well-informed clinical advice and monitoring the progress of patients with complex mental health issues. To develop a superior CDSS, it's necessary to create an integrated ecosystem that effectively serves both patients and the wider healthcare team, including social workers, nurses, doctors, and domain experts. A key feature of such a system is its ability to assess risks, enabling it to offer personalized recommendations based on individual patient risks, a focus of this project.

The concept of a "digital clinic", supported by medical professionals, emerges as a particularly innovative model. These clinics provide patients with access to expert medical advice, regardless of geographical constraints. Additionally, partnerships with online pharmacies and similar platforms could simplify the healthcare process for patients, enhancing the role of CDSS as not just a tool, but a comprehensive healthcare solution that works in tandem with existing medical facilities.

Recent developments have brought attention to platforms such as the Innowell from the University of Sydney and the Risk Assessment and Management Platform from the University of British Columbia (214,215). These platforms go beyond mere technological advancements; they are crucial tools that improve the quality of care. They ensure that patients receive timely and accurate interventions within a cohesive healthcare environment.
3.4.5 Navigating Limitations and Ensuring Ethical Deployment

The introduction of Clinical Decision Support Systems (CDSS) offers numerous advantages, yet it also presents certain challenges. A significant concern is alert fatigue, and variations in computer literacy among physicians and patients can hinder the full utilization of CDSS capabilities (303). Ongoing maintenance and updates of these systems are crucial to maintain the accuracy and relevance of the advice they provide.

Additionally, the ethical issues arising from AI-driven biases in CDSS require careful consideration. The non-knowledge-based versions, despite their benefits, often suffer from a lack of transparency. The complex algorithms that drive their predictions and decisions can be difficult for users to understand, which can lead to diagnostic inaccuracies due to a failure to recognize subtle details or prepare for rare events. This opacity is a consequence of the sophisticated models employed, posing a challenge in fully grasping the logic behind specific prognoses and choices.

3.5 Conclusion

The incorporation of Clinical Decision Support Systems (CDSS) in healthcare, especially in addressing addiction and concurrent disorders, marks a significant development. Our findings show that CDSS has improved clinical decision-making and adherence to guidelines, but its immediate effect on addiction outcomes is relatively limited in statistical terms. As healthcare providers become more acquainted and confident with CDSS, its use and effectiveness are expected to enhance, potentially leading to better clinical outcomes. There is a need for additional research to assess the effectiveness of CDSS in specific treatment approaches, and to
explore how healthcare providers' acceptance of CDSS affects their implementation of its treatment recommendations.

Our study also observes a shift in CDSS methodologies from traditional knowledge-based to more contemporary data-driven models. This shift corresponds with advancements in computational technology and big data analytics, which could impact personalized medicine. However, ethical considerations and accurate application of these evolving systems are critical areas that need ongoing attention. The study emphasizes the importance of making CDSS accessible to marginalized groups, calling for a holistic and inclusive approach that involves a diverse healthcare team.

In summary, CDSS holds promise for enhancing healthcare decision-making and patient outcomes, but its successful and ethical integration into healthcare requires a well-rounded, multi-dimensional strategy. Ongoing research is crucial to address the intricacies of CDSS implementation and to ensure its effective and fair use in various healthcare contexts.

3.5.1 Study Limitation

Our systematic review encountered a few limitations. The exclusive inclusion of English-language studies from specific databases could result in selection bias. Furthermore, the heterogeneity in the study populations and outcomes used in different studies poses a challenge for making quantitative comparisons. Additionally, this variability, coupled with occasional lapses in the reporting of CDSS evaluation, hinders our capacity to compare findings consistently and directly across the studies.
### 3.6 Risk of Bias

#### 3.6.1 Randomized Control Trials

Table 4 Risk of bias assessment for Randomized Control Trials, Risk Of Bias 2 (1) for Clinical Decision Support in Addiction and Concurrent Disorders: A Systematic Review and Meta-Analysis

<table>
<thead>
<tr>
<th>Study ID</th>
<th>Experimental</th>
<th>Comparator</th>
<th>Outcome</th>
<th>Weight</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
<th>D5</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rollman et al., 2002</td>
<td>CDSS use: active care, pass</td>
<td>Usual care</td>
<td>3 month and 6 month clin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Kriek et al., 2013</td>
<td>Electronic decision aid</td>
<td>Usual care</td>
<td>Electronic decision aids</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Fletcher et al., 2021</td>
<td>Prognosis-matched care</td>
<td>Usual care plus attention coi Prognosis-matched care</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Minian et al., 2020</td>
<td>Delivery and acceptance of Delivery and acceptance</td>
<td>Delivery and acceptance of f</td>
<td>Delivery and acceptance</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Bergman &amp; Flors, 2008</td>
<td>Diagnostic outcome using</td>
<td>Diagnostic outcome using</td>
<td>No significant difference</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Bernstein et al., 2017</td>
<td>Best practice alert offersto</td>
<td>BPA silently fires with no adv Physicians more likely to</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Medow et al., 2001</td>
<td>Computerized DSA OR writ</td>
<td>Assessing scenario without a Written DSA Improved d</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Perestelo-Perez et al., 2017</td>
<td>Decision aid</td>
<td>Usual care</td>
<td>Decision aid improved kr</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Balesbiteri et al., 2020</td>
<td>CCDSS and telemedicine to Treatment as usual</td>
<td>CCDSS and telemedicine</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Lutz et al., 2022</td>
<td>CDSS</td>
<td>No CDSS</td>
<td>Differential effect size of</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Robinson et al., 2018</td>
<td>NAVIGATE treatment</td>
<td>Clinician-choice community NAVIGATE participants h</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>LeBlanc et al., 2015</td>
<td>Treatment of depression wi Treatment of depression wi</td>
<td>Decision-making quality</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Melnick et al., 2022</td>
<td>CDSS</td>
<td>Usual care</td>
<td>Rate of initiation of bupr</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Zastrozhin et al., 2018</td>
<td>Treatment using pharmoge Treatment as usual</td>
<td>Treatment as usual</td>
<td>Side effects</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Rosson et al., 2021</td>
<td>OUD-CDSS use</td>
<td>Treatment as usual</td>
<td>Accuracy of the OUD-CD1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Kuriyama et al., 2009</td>
<td>MDD treatment using CDSS Treatment as usual</td>
<td>Treatment as usual</td>
<td>Depression symptom sev</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Co et al., 2010</td>
<td>Electronic health record de Treatment as usual</td>
<td>Treatment as usual</td>
<td>Proportion of children w</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Harrison et al., 2023</td>
<td>CDSS</td>
<td>Treatment as usual</td>
<td>Depressive symptoms, fe</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td>Dobscha et al., 2006</td>
<td>Depression decision suppo Treatment as usual</td>
<td>Change in depression sco</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low risk</td>
</tr>
</tbody>
</table>

- D1: Randomisation process
- D2: Deviations from the intended interventions
- D3: Missing outcome data
- D4: Measurement of the outcome
- D5: Selection of the reported result

85
Figure 16 Risk of Bias Assessment in Intention-to-Treat Analysis: ROB2 Visualization for Clinical Decision Support in Addiction and Concurrent Disorders: A Systematic Review and Meta-Analysis

Table 5 Risk of bias assessment for COHORT studies, Modified Newcastle Ottawa Quality Assessment Scale (196) for Clinical Decision Support in Addiction and Concurrent Disorders: A Systematic Review and Meta-Analysis

<table>
<thead>
<tr>
<th>Author, year</th>
<th>Representativeness of the exposed cohort</th>
<th>Selection of the non-exposed cohort</th>
<th>Ascertaiment of exposure</th>
<th>Demonstration that outcome of interest was not present at the start of the study</th>
<th>Adjusts for most important risk factor</th>
<th>Adjusts for other risk factor(s)</th>
<th>Assessmen of outcome</th>
<th>Was follow-up long enough for outcome</th>
<th>Adequacy of follow-up of cohorts</th>
<th>Score (/9)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
At least somewhat representative of average ___ in the community | Drawn from same community as exposed cohort | Structured interview or secure record | Yes | Yes | Yes | Record linkage or independent blind assessment | Yes | Complete follow up or subjects lost unlikely to introduce bias
---|---|---|---|---|---|---|---|---
Brown et al., 2015 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 4
Wai et al., 2019 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 9

Table 6 Risk of bias assessment for randomized controlled trials or case control studies, Modified Newcastle Ottawa Quality Assessment Scale (196) for Clinical Decision Support in Addiction and Concurrent Disorders: A Systematic Review and Meta-Analysis

<table>
<thead>
<tr>
<th>Author, year</th>
<th>Modified Newcastle Ottawa Quality Assessment Scale for RANDOMIZED CONTROLLED TRIALS or CASE CONTROL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SELECTION (max 4)</td>
</tr>
<tr>
<td>Is the case definition adequate?</td>
<td>Representativeness of the cases</td>
</tr>
<tr>
<td>Yes WITH independent validation</td>
<td>Consecutive or obv. representative</td>
</tr>
</tbody>
</table>

87
<table>
<thead>
<tr>
<th>Study</th>
<th>Year</th>
<th>Case</th>
<th>Control</th>
<th>Sample</th>
<th>Setting</th>
<th>Exposure</th>
<th>Outcome</th>
<th>x</th>
<th>Control</th>
<th>Outcome</th>
<th>x</th>
<th>Case</th>
<th>Control</th>
<th>Outcome</th>
<th>x</th>
<th>Case</th>
<th>Control</th>
<th>Outcome</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trivedi et al., 2004</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ma et al., 2017</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malhotra et al., 2019</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rollman et al., 2002</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kriek et al., 2013</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fletcher et al., 2021</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bergman &amp; Fors, 2005</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minian et al., 2020</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bergman &amp; Fors, 2008</td>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bernstein et al., 2017</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medow et al., 2001</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perestello-Perez et al., 2017</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balestrieri et al., 2020</td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Mooney et al., 2020  
Melnick et al., 2020  
Lutz et al., 2022  
Robinson et al., 2018  
LeBlanc et al., 2015  
Kingslake et al., 2017  
Zastrozhin et al., 2018  

<table>
<thead>
<tr>
<th>Author, year</th>
<th>Representativeness of the sample</th>
<th>Sample size</th>
<th>Non-respondents</th>
<th>Ascertainment of exposure (max 2)</th>
<th>Controls for most important risk factor</th>
<th>Controls for other risk factor(s)</th>
<th>Assessment of outcome (max 2)</th>
<th>Statistical test</th>
<th>Score (/10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 7 Risk of bias assessment for cross-sectional studies, Modified Newcastle Ottawa Quality Assessment Scale (196) for Clinical Decision Support in Addiction and Concurrent Disorders: A Systematic Review and Meta-Analysis
At least somewhat representative of average ___ in the community

Justified and satisfactory

Satisfactory response rate and respondents and non-respondent s established to be comparable

2: Validated measurement tool
1: Non-validated measurement tool, but tool is described

Yes

Yes

2: Record linkage or independent blind assessment
1: Self report

Test is described, appropriate, and association is presented with confidence intervals and p value

Ally & Stallman., 2016
Etter et al., 2018

<table>
<thead>
<tr>
<th>Author, year</th>
<th>Was there a clear statement of the aims of the research?</th>
<th>Is a qualitative methodology appropriate?</th>
<th>Was the research design appropriate to address the aims of the research?</th>
<th>Was the recruitment strategy appropriate to the aims of the research?</th>
<th>Was the data collected in a way that addressed the research issue?</th>
<th>Has the relationship between researcher and participant been adequately considered?</th>
<th>Have ethical issues been taken into consideration?</th>
<th>Was the data analysis sufficient and rigorous?</th>
<th>Is there a clear statement of findings?</th>
<th>How valuable is the research?</th>
<th>Score (/10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooper et al., 2006</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 8 Risk of bias assessment, Critical Appraisal Skills Programme (304,305) for Clinical Decision Support in Addiction and Concurrent Disorders: A Systematic Review and Meta-Analysis
<table>
<thead>
<tr>
<th>Reference</th>
<th>Year</th>
<th>No. 1</th>
<th>No. 2</th>
<th>No. 3</th>
<th>No. 4</th>
<th>No. 5</th>
<th>No. 6</th>
<th>No. 7</th>
<th>No. 8</th>
<th>No. 9</th>
<th>No. 10</th>
<th>No. 11</th>
<th>No. 12</th>
<th>No. 13</th>
<th>No. 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kleinerman et al., 2022</td>
<td>2022</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qassim et al., 2022</td>
<td>2022</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brenton et al., 2018</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Henshall et al., 2017</td>
<td>2017</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohen, 2015</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trivedi et al., 2009</td>
<td>2009</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zisman-Ilani et al., 2018</td>
<td>2018</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haroz et al., 2021</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goorden et al., 2017</td>
<td>2017</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krugten et al., 2021</td>
<td>2021</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berrouiguet et al., 2016</td>
<td>2016</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berrouiguet et al., 2017</td>
<td>2017</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Popescu et al., 2021</td>
<td>2021</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tanguay-Sela et al., 2021</td>
<td>2021</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Holland et al., 2020</td>
<td>2020</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pierce et al., 2022</td>
<td>2022</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Henshall et al., 2019</td>
<td>2019</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bonfils et al., 2018</td>
<td>2018</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Torrey et al., 2020</td>
<td>2020</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Year</td>
<td>Study Type</td>
<td>Rating</td>
<td>Impact</td>
<td>Evidence Level</td>
<td>Quality Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td>------</td>
<td>------------</td>
<td>--------</td>
<td>--------</td>
<td>----------------</td>
<td>---------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buckingham et al., 2015</td>
<td>2015</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>McNeely et al., 2022</td>
<td>2022</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goodspeed et al., 2019</td>
<td>2019</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ferns, 1995</td>
<td>1995</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rast et al., 2020</td>
<td>2020</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trivedi &amp; Daly, 2007</td>
<td>2007</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whelan et al., 2015</td>
<td>2015</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bronzino et al., 1989</td>
<td>1989</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aalsma et al., 2018</td>
<td>2018</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beaudenquin et al., 2020</td>
<td>2020</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chu et al., 2016</td>
<td>2016</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clausen et al., 2020</td>
<td>2020</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gibbons et al., 2013</td>
<td>2013</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tasma et al., 2018</td>
<td>2018</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roebroek et al., 2022</td>
<td>2022</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lutz et al., 2019</td>
<td>2019</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clausen et al., 2023</td>
<td>2023</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Melnick et al., 2019</td>
<td>2019</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gill et al., 2012</td>
<td>2012</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

92
Chapter 4: Data strategies and pilot model for the prediction of general and fatal overdose

4.1 Introduction

4.1.1 Background

As described in chapter 1, BC is amidst an drug toxicity crisis, which has taken 14,769 to date (19). Chapter 2’s systematic review and meta-analysis showed evidence to suggest machine learning models such as logistic regression, random forest, boosting algorithms and deep learning are viable methods to predict upon opioid-use related outcomes such as overdose. Chapter 3’s systematic review and meta-analysis showed how predictive models could be more effective when provided alongside treatment. Hence, machine learning models were created to predict the general and fatal overdose.

4.1.2 Objectives and hypotheses

This chapter describes the utilization of the British Columbia Provincial Overdose Cohort (BCODC) collected by the British Columbia Centre for Disease control to a dataset to predict upon general and fatal overdose.

4.2 Methods

4.2.1 Data access

The BCODC was accesses through Population Data BC through a secure research environment (SRE) (306).
Access to data provided by the Data Stewards is subject to approval but can be requested for research projects through the Data Stewards or their designated service providers. The following data sets were used in this study: British Columbia Overdose Cohort, including healthcare service billings from the Medical Service Plan (MSP), hospitalization records from the Discharge Abstract Database (DAD), prescription details from PharmaNet, outpatient care data from the National Ambulatory Care Reporting System (NACRS), demographic information from the Client Registry, emergency response records from BC Emergency Health Services (BC EHS), mortality data from the BC Coroner’s Service (BCCS) and the Vital Statistics Deaths database, inquiries made to the Drug and Poison Information Centre (DPIC), chronic disease cases from the Chronic Disease Registry, mental health service records from the Mental Health Data Warehouse (MHDW), and socioeconomic information from the Ministry of Social Development and Poverty Reduction (SDPR).

You can find further information regarding these data sets by visiting the PopData project webpage at: https://www.popdata.bc.ca/data/health/ODC. All inferences, opinions, and conclusions drawn in this publication are those of the author(s), and do not reflect the opinions or policies of the Data Steward(s).

4.2.2 Research Environment

Research was done through Population Data BC’s (PopData) secure research environment (SRE) (306). The SRE, offered by PopData, serves as a specialized service for researchers accessing data via PopData (306). It acts as a centralized hub for data access, processing, and secure storage, including data extract backups. Researchers are provided with complimentary access to data analysis software within the SRE.
Security in the SRE is paramount, with access restricted to a secure private cloud. This access is safeguarded through an encrypted Virtual Private Network (VPN), reinforced by a firewall, and further secured using YubiKey® token authentication (306). PopData's solution offers researchers a reliable and secure method for managing data extracts, ensuring compliance with Data Stewards' security standards and significantly reducing the risk of unauthorized access to sensitive personal data.

In the Secure Research Environment (SRE), data analysis was conducted using RStudio version 1.3.1093, paired with R version 4.0.5 (released on 2021-03-21), on an X86_64-w64-mingw32/x64 (64-bit) platform. The system was operated under Windows 10 x64 (build 19045).

4.2.3 Data Wrangling

For data wrangling, some basic packages were utilized which was included in the tidyverse packages version 2.0.0 (307). These packages included dplyr 1.1.2, readr 2.1.3, forcats 1.0.0, stringr 1.5.0, ggplot2 3.4.2, tibble 3.2.1, lubridate 1.9.2, tidyr 1.3.0, purr 1.0.1.

4.2.3.1 Overdose Data

To construct a predictive machine learning model, it is essential to develop a comprehensive dataset derived from the British Columbia Overdose Cohort (BCODC). Firstly, the dataset Overdose Cases was utilized. Overdoses were identified based on overdose cases in people with at least one general or fatal overdose episode between January 1st, 2015 and December 31st, 2017 (43). As mentioned in chapter 1.5, Overdose incidents were identified from BC Emergency Health Services, Drug and Poison Information Centre, BC Coroner's Service, and
case reports from Interior, Vancouver Island, and Northern Health Authorities' EDs, along with national care reporting, hospital discharge data, medical billing, and mortality records (43). As a result, an overdose case dataset was built which identified by ministry of health identification, individuals who have had at least one general or fatal overdose episode between January 1st, 2015 and December 31st, 2019. The utilization of this dataset was core to this project.

4.2.3.2 Pilot model

In the initial phase of developing a machine learning model to predict fatal overdoses, the focus was on the Coroner's Service dataset, which was anticipated to be a crucial resource. Access to the cohort dataset led to an initial concentration on data from the Coroner's Service due to its comprehensive reporting of fatalities.

The selection of the Coroner's Service as the initial data source was based on its comprehensive documentation of deceased individuals, which suggested it as a potentially ideal foundation. However, the search for comparable data for the control group—individuals who survived overdoses—presented challenges. Examination of various datasets, including the BC Emergency Health Services (BCEHS), the Drug and Poison Information Centre (DPIC), Enhanced Emergency Department Records (EDD), and the National Ambulatory Care Reporting System (NACRS), uncovered inconsistencies in the data collection process. Detailed descriptions of these databases can be found in Appendix C.

One notable gap was the lack of fentanyl exposure data for general overdose survivors, which contrasted with the detailed records of fentanyl-related fatalities from the Coroner's Service. The datasets employed in this initial modeling effort are delineated in Figure 17. A significant realization emerged from this discrepancy: the other 13 datasets lacked comparative
data for overdose survivors. The original intent was to create a contrast between fatal overdose cases from the Coroner's Service and general overdose controls. However, this approach led to a significant class imbalance, heavily favoring fatal cases at a ratio of 10:1, with a mere 600 general cases included. Consequently, the pilot model developed under these constraints yielded suboptimal results.

Faced with this imbalance and data inconsistency, it became necessary to re-evaluate the approach and search for a more inclusive dataset that properly encompassed both fatal and general overdose instances. The initial inclusion of the Chronic Disease Registry (CDR), with its standardized International Classification of Diseases (ICD) codes for creating variables, led to an essential realization: leveraging these ICD codes might lead to a more equitable dataset for predictive modeling.

Figure 17 Data Sources Overview for the Pilot Predictive Model
4.2.3.3 Dataset selection and organization

The study prominently featured five key datasets: Overdose Case, PharmaNet, Discharge Abstract Database, Medical Services Plan, and Chronic Disease Registry, all of which are illustrated in Figure 18. Each dataset was chosen for its comprehensive coverage of different aspects of healthcare data. The Overdose Case Dataset is central to the analysis, providing information on the overdose population including data on fatal overdoses and the frequency of overdoses per individual. Pharma Net, encompassing data from all pharmacies in BC, offers a complete picture of prescription patterns, crucial for understanding medication-related overdoses. The Discharge Abstract Database contributes hospital-level data, including information about patient hospital stays, diagnoses, and treatments, which is vital for understanding the acute care aspect of overdose cases. The Medical Services Plan, representing provincial insurance data, captures a broad spectrum of healthcare encounters, including those not directly related to hospitalizations but pertinent to overdose incidents. Lastly, the Chronic Disease Registry provides insights into the prevalence of chronic health conditions among individuals experiencing overdoses, adding context to the overall health status of the affected population. This combination of datasets enables the creation of a comprehensive and multi-dimensional view of healthcare utilization surrounding overdose events. Detailed descriptions of these databases are available in Appendix C. A summary of each of these tables is also provided in Appendix C.
derived variables

As previously mentioned, to address the encountered data imbalance and inconsistency, a methodology like that used in the Chronic Disease Registry was adopted. This approach involved leveraging data from hospital records, insurance claims, and pharmacy dispensations to derive variables. This strategy aimed to capture a broader population, thereby furnishing this project with a dataset that was not only more balanced but also more comprehensive. The analysis incorporated data that detailed the history of diagnoses spanning the years 2015 to 2019.

The Medical Services Plan utilized the International Classification of Diseases, 9th Edition (ICD-9) coding system, while the Discharge Abstract Database employed the 10th Edition (ICD-10) of the same system (185,186,308,309). For identifying Opioid Agonist
Therapy (OAT) in the PharmaNet data, Drug Identification Numbers/Product Identification Numbers (DIN/PIN) were used (310).

In line with the algorithms for drug misuse established by Janjua et al., this project applied a criterion of either one medical visit (coded with ICD-9) or one hospitalization (coded with ICD-10) to identify specific categorical features (311). This approach facilitated the derivation of variables vital to predicting outcomes related to general and fatal overdose. These variables can be seen in Table 9. All ICD-9, ICD-10 and DIN/PIN codes utilized for these algorithms can be seen in Appendix D.
<table>
<thead>
<tr>
<th>Variables</th>
<th>Data source</th>
<th>Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOH Study ID</td>
<td>Ministry of Health</td>
<td>N/A</td>
</tr>
<tr>
<td>Opioid Agonist Treatment</td>
<td>Pharmanet</td>
<td>1+ DIN/PIN within 5 years</td>
</tr>
<tr>
<td>Rheumatoid Arthritis</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Parkinsonism</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Osteoarthritis</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Multiple Sclerosis</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Mood Anxiety</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Ischemic Stroke</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Ischemic Heart Disease</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Hypertension</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Hospital Transient Ischemic Attack</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Hospital Stroke</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Heart Failure</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Hemorrhagic Stroke</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Epilepsy</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Depression</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Chronic Obstructive Pulmonary Disease</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Chronic Kidney Disease</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Asthma</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Angina</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.1</td>
</tr>
<tr>
<td>Acute Myocardial Infarction</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.2</td>
</tr>
<tr>
<td>Alzheimer Dementia</td>
<td>Chronic Disease Registry</td>
<td>check appendix C.2</td>
</tr>
<tr>
<td>Tobacco Use</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 years</td>
</tr>
<tr>
<td>Tissue Infection</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 years</td>
</tr>
<tr>
<td>Condition</td>
<td>Database</td>
<td>Timeframe</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>-----------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Substance Related Disorders</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Stimulant Use</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Sepsis</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Sedative Hypnotic Use</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Psychotic Disorders</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Polysubstance Use</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Personality Disorders</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Other Psychoactive Drug Use</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Osteomyelitis</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Opioid Use</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Neurotic Related Disorders</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Neurocognitive Disorders</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Multiple Mental Illness</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Mood Disorders</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Intellectual Disability</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Hallucinogen Use</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Endocarditis</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Early Onset Disorders</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Developmental Disorders</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Cocaine Use</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Cannabinoid Use</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Behavioural Psychological Disturbances</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Alcohol Use</td>
<td>Medical Service Plan/Discharge Abstract Database</td>
<td>1+ H or 2+ P within 5 year</td>
</tr>
<tr>
<td>Total Overdose Number</td>
<td>Overdose Case</td>
<td>Check section 4.3.2.1</td>
</tr>
<tr>
<td>Fatal Overdose Case</td>
<td>Overdose Case</td>
<td>Check section 4.3.2.1</td>
</tr>
<tr>
<td>general Overdose Case</td>
<td>Overdose Case</td>
<td>Check section 4.3.2.1</td>
</tr>
</tbody>
</table>
Abbreviations: H, hospitalization; P, physician/MSP visit; Rx, drug prescription. Detailed code of derived variables from MSP/DAD can be seen in Appendix C.
4.2.4 Results

4.2.4.1 Control data

The control data set was formulated by joining PharmaNet data and ICD-derived variables onto CDR data. After excluding the overdose population, this resulted in a data set comprising 608,182 unique Ministry of Health IDs representing the control group.

4.2.4.2 Overdose data

The overdose data set was created by integrating PharmaNet data and ICD-derived variables with CDR data specific to the overdose population. This process yielded a total of 36,679 unique Ministry of Health IDs. Among these, 30,253 individuals had general overdoses, while 6,426 experienced fatal overdoses.

4.2.5 Discussion

Results from chapter 4, which created datasets, were utilized in chapter 5 and 6 for machine learning predictive modeling, and inferenced based modelling.
Chapter 5: Modeling for overdose: utilizing machine learning in predicting fatal and general overdose

5.1 Introduction

This chapter is the main work of this dissertation, which describe the utilization of machine learning models to predict fatal and general overdose.
Modeling for overdose: utilizing machine learning in predicting fatal overdose.

5.2 Objective

This chapter describes the utilization of the BC Provincial Overdose Cohort (BCODC) collected by the BC Centre for Disease control to create multiple machine learning models predicting upon general overdose.

5.3 Methods

In this study, 'controls' refer to individuals who experienced general overdoses, while 'cases' denote individuals who experienced fatal overdoses. All data analysis was conducted using RStudio (version 1.3.1093). Details of the coding packages/libraries employed, as well as the raw code, are provided in Appendix D. The analysis incorporated data that detailed the history of diagnoses spanning the years 2015 to 2019.

5.3.1 Dataset

Table 10 Variables included in predictive modelling for fatal overdose

<table>
<thead>
<tr>
<th>Variables</th>
<th>Missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOH Study ID</td>
<td>0</td>
</tr>
<tr>
<td>Opioid Agonist Treatment</td>
<td>3306</td>
</tr>
<tr>
<td>Rheumatoid Arthritis</td>
<td>5175</td>
</tr>
<tr>
<td>Parkinsonism</td>
<td>5175</td>
</tr>
<tr>
<td>Osteoarthritis</td>
<td>5175</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>5175</td>
</tr>
<tr>
<td>Multiple Sclerosis</td>
<td>5175</td>
</tr>
<tr>
<td>Mood Anxiety</td>
<td>5175</td>
</tr>
<tr>
<td>Ischemic Stroke</td>
<td>5175</td>
</tr>
</tbody>
</table>
Ischemic Heart Disease
Hypertension
Hospital Transient Ischemic Attack
Hospital Stroke
Heart Failure
Hemorrhagic Stroke
Epilepsy
Diabetes
Depression
Chronic Obstructive Pulmonary Disease
Chronic Kidney Disease
Asthma
Angina
Acute Myocardial Infarction
Alzheimer Dementia
Tobacco Use
Tissue Infection
Substance Related Disorders
Stimulant Use
Sepsis
Sedative Hypnotic Use
Psychotic Disorders
Polysubstance Use
Personality Disorders
Other Psychoactive Drug Use
Osteomyelitis
Opioid Use
Neurotic Related Disorders
Neurocognitive Disorders
Multiple Mental Illness
Mood Disorders
Intellectual Disability
Hallucinogen Use
Endocarditis
Early Onset Disorders
Developmental Disorders
Cocaine Use
Cannabinoid Use
Behavioural Psychological Disturbances
Alcohol Use
5.3.2 Data Preprocessing

5.3.2.1 Sample size

This study's sample comprised 30,253 general overdose cases classified as controls, and 6,426 fatal overdose cases categorized as the case group, resulting in a total sample size of 36,679. This composition reflects the distribution of general and fatal overdose cases in the study.

5.3.2.2 Data splitting

The train test split for machine learning was done utilizing the Classification And REgression Training (Caret), 6.0-94 package (312). Data was partitioned into 70% training data and 30% testing data. After partitioning, the train set had 25676 observations, with controls being 21212 and cases being 4464. The test set had 11003 observations, with controls being 9041 and cases being 1962. All data preprocessing was done after the split, to adhere to the golden rule of machine learning (313,314).

5.3.2.3 Missing Data

Cases have 337 missing rows for data from MSP and DAD derived variables. 939 missing rows for data from CDR. 684 missing rows for OAT from Pharmanet. A total of 25676 sample.

Controls have 1398 missing rows for data from MSP and DAD derived variables. 4236 missing rows for data from CDR. 2622 missing rows for OAT from Pharmanet.
After the train-test split, the train data have 1190 missing rows for data from MSP and DAD derived variables. 3611 missing rows for data from CDR. 2306 missing rows for OAT from Pharmanet.

After the train-test split, the train data have 545 missing rows for data from MSP and DAD derived variables. 1564 missing rows for data from CDR. 1000 missing rows for OAT from Pharmanet.

The Multiple Imputation by Chained Equations (MICE) package, version 3.15.0, in RStudio was utilized for handling missing binary outcome data (315). The adopted methodology was Classification and Regression Trees (CART) from the MICE package, which has demonstrated substantial improvements in imputation over traditional parametric methods due to its recursive partitioning approach (316). The number of imputations performed was set at 15 which was the most a column has for percentage of missing data, a quantity supported by the literature as robust (317–320).

### 5.3.2.4 Resampling

The train set had 25676 observations, with controls being 21212 and cases being 4464. The data comprised of of 82.6% controls and 17.4% cases.

The training set's class balance was achieved through the ROSE package, version 0.0-4, which applies random over and under sampling techniques to equalize the representation of different classes (321). This approach is employed during the training phase to prevent model bias toward more frequently occurring classes. In contrast, the test set was not subjected to class balancing, aligning with the objective of assessing the model's performance on real-world,
potentially imbalanced data (322). Random over-sampling, despite its simplicity, competes effectively with more intricate methods in preparing the data for training robust models (323).

In this study, a range of resampling techniques were employed to balance the training dataset, including random undersampling, random oversampling, and a combination of both strategies. The undersampling method adjusted the ratio of controls (individuals with general overdoses) to cases (individuals with fatal overdoses) to both 3:1 and 1:1. Additionally, a combined approach of over and under sampling was utilized to further balance the dataset. The oversampling strategy was implemented to achieve an equal ratio of cases to controls.

The test data was maintained in its original state to accurately reflect the real-world prevalence of fatal overdoses within the overall overdose population. This approach ensures that the predictive modeling closely mirrors the actual conditions and challenges encountered in addiction psychiatry, particularly in the context of overdose risk prediction.

5.3.2.5 Machine learning

Several machine learning models were employed using packages such as Caret and Tidy Models. Specifically, XGBoost and Random Forest models were utilized to assess the performance across various class-balanced samples from the Caret Package (caret, 6.0-94) (312). It is important to note that no cross-validation was performed in this initial testing phase for these models. These models were then tested and compared to see which balanced dataset was performing the best.

In a more comprehensive phase of the study, a range of models including XGBoost, Random Forest, Adaptive Boosting with Bagging (AdaBag), Support Vector Machine (SVM), Gradient Boosting Machine (GBM), Naïve Bayes, and K-Nearest Neighbor (KNN) were trained,
underwent 10-fold cross-validation, and were tested from the Caret Package (caret, 6.0-94) (312). This approach provided a robust evaluation of the models' performance. After cross validation, these models were then tested and compared to see which was performing the best.

Furthermore, using the Tidy Models package, additional models such as XGBoost, Random Forest, Generalized Linear Model (GLM), and K-Nearest Neighbors were applied, all undergoing the same training, 10-fold cross-validation, and testing process from the tidymodels package (tidymodels, 0.1.1.) (324). After cross validation, these models were then tested and compared to see which was performing the best.

Lastly, deep learning techniques were explored using the Keras package, implementing a neural network with 5 hidden layers (325,326). This allowed for a comparison between traditional machine learning approaches and more complex, deep learning models. The deep learning models were then tested with the test data.

### 5.3.2.5.1 Hyper parameter tuning

In this study, the machine learning models were developed using the Caret package in R, with hyperparameter tuning achieved through Caret's train control functionality. The XGBoost model, a boosting method, sequentially adds predictors (trees), where each tree aims to correct the errors of its predecessor. The 'nrounds' parameter in XGBoost dictates the number of boosting rounds, essentially determining how many trees are added. The balance between too many trees leading to overfitting and too few resulting in an overly general model was carefully considered. To mitigate the risk of overfitting, a decision was made to use 300 trees. This was also the same methodology utilized for Adabag with 300 trees. Adabag is a model that implements two of the most popular ensemble learning algorithms: Adaptive
Boosting (AdaBoost) and Bootstrap Aggregating (Bagging). Ensemble learning is a technique where multiple models (often called "weak learners") are trained and combined to solve a particular computational intelligence problem, often resulting in better performance than any single model could achieve.

Random Forest, implemented via the ranger package in Caret, combines bagging (bootstrap aggregating) with the random subspace method. Bagging involves using subsets of observations, rather than the entire dataset, to grow each tree, while the random subspace method selects a subset of features for each tree. Given the potentially large dataset size but a limited number of trees, there's a chance that some observations might contribute only once or not at all to the prediction. Similarly, with numerous predictors and fewer trees, some features might not be included in any tree's subspace. To address this, 500 trees were chosen as the default setting in the ranger package, considering that Random Forest is less prone to overfitting. The 'mtry' parameter, representing the number of features considered at each split, was set to 7 (76), following the standard practice of using the square root of the total number of predictors for classification tasks. Furthermore, the Gini impurity was employed as the default criterion for splitting.

The logistic regression that was utilized has no regularization. A set number of regularization strengths was utilized in this study. The values listed (0.001, 0.01, 0.1, 1, 10, 100, 1000) represent examples of these strengths (327). Regularization strength plays a crucial role in model fitting: a smaller value indicates stronger regularization, applying more penalty to large coefficients, while a larger value denotes weaker regularization, imposing less penalty. This approach is integral to optimizing the balance between model complexity and generalization capability.
For the Support Vector Machine, Gradient Boosting Machine, Naïve Bayes, and K-Nearest Neighbor models, the default hyperparameters were used.

5.3.2.5.2 Evaluation metrics

To evaluate each model, 10-fold cross validation was done (328). This technique involves dividing the training data into 10 distinct subsets. Each subset is used as a test set while the remaining nine subsets collectively serve as the training set. This process is repeated ten times, ensuring that each subset is used exactly once as the test set. This method provides a comprehensive evaluation, testing the model's effectiveness across various segments of the training data.

To test each model, the test data was segregated before any preprocessing activities were conducted on the dataset (329). This preprocessing involved handling missing data and balancing the classes. Following this, predictions were made using the test data, and a confusion matrix was generated using the Caret package (caret, version 6.0-94) (312). This confusion matrix provided several key metrics: accuracy, sensitivity, specificity, positive predictive value, and negative predictive value (330). Additionally, the area under the receiver operating characteristic (AUROC) curve was calculated using the RocR package (ROCR, version 1.0-11), offering further insight into model performance (331).

Accuracy gauges the model's overall correctness, while Sensitivity and Specificity assess its ability to accurately identify actual cases and controls, respectively. PPV indicates the likelihood of true positive predictions, and NPV reflects the reliability in identifying non-risk individuals. The AUROC measures the model's capacity to distinguish between cases and controls. As highlighted in Chapter 2, these metrics are among the most frequently utilized in
literature for assessing the efficacy of machine learning models predicting opioid related outcomes.

5.3.2.5.3 Confidence intervals

To make confidence intervals from AUC scores, Wilson’s score was utilized (113).

\[ \text{CI}_{\text{lower}}, \text{CI}_{\text{upper}} = \left( \frac{1}{1+\frac{z^2}{N}} \right) \left( p + \frac{z^2}{2n} \pm z \sqrt{\left( \frac{p(1-p)}{n} + \frac{z^2}{4n^2} \right)} \right) \]

5.4 Results

5.4.1 Resampling results

Due to the significantly lower number of fatal overdose cases (6,426) compared to controls, various resampling strategies were utilized. The implementation of resampling strategies resulted in different ratios of controls to cases. In the under sampling approach, two distinct ratios were achieved: a 3:1 ratio resulting in 12,039 controls and 4,464 cases, and a 1:1 ratio with 4,464 controls and cases each. The combined method of over and under sampling led to a balanced count of 12,876 controls and 12,800 cases. Through the oversampling strategy, a 1:1 ratio was achieved with 21,212 controls and 21,018 cases. These strategies successfully addressed the imbalance inherent in the original dataset, enhancing the robustness and reliability of the analysis. The equitable representation of cases and controls achieved through these methods was vital for accurate model training and subsequent analysis.

5.4.2 Machine learning results

Table 11 Evaluation of class balanced sample when predicting fatal overdose
<table>
<thead>
<tr>
<th>Package</th>
<th>Model type</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive Predictive Value</th>
<th>Negative Predictive Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caret</td>
<td>XGBoost Orginal</td>
<td>81.34%</td>
<td>4.49%</td>
<td>98.02%</td>
<td>32.96%</td>
<td>82.54%</td>
</tr>
<tr>
<td>Caret</td>
<td>Random Forest Original</td>
<td>82.25%</td>
<td>0.51%</td>
<td>99.99%</td>
<td>90.91%</td>
<td>82.24%</td>
</tr>
<tr>
<td>Caret</td>
<td>XGBoost Under 3:1</td>
<td>78.71%</td>
<td>15.09%</td>
<td>92.51%</td>
<td>30.42%</td>
<td>83.39%</td>
</tr>
<tr>
<td>Caret</td>
<td>Random Forest Under 3:1</td>
<td>82.24%</td>
<td>1.53%</td>
<td>99.76%</td>
<td>57.69%</td>
<td>82.36%</td>
</tr>
<tr>
<td>Caret</td>
<td>XGBoost Under 1:1</td>
<td>58.39%</td>
<td>60.30%</td>
<td>58.35%</td>
<td>23.90%</td>
<td>87.15%</td>
</tr>
<tr>
<td>Caret</td>
<td>Random Forest Under 1:1</td>
<td>61.13%</td>
<td>59.94%</td>
<td>25.20%</td>
<td>87.59%</td>
<td>17.83%</td>
</tr>
<tr>
<td>Caret</td>
<td>XGBoost Both 1:1</td>
<td>63.87%</td>
<td>53.77%</td>
<td>66.07%</td>
<td>25.59%</td>
<td>86.82%</td>
</tr>
<tr>
<td>Caret</td>
<td>Random Forest Both 1:1</td>
<td>66.42%</td>
<td>48.01%</td>
<td>70.41%</td>
<td>26.04%</td>
<td>86.19%</td>
</tr>
<tr>
<td>Caret</td>
<td>XGBoost Over 1:1</td>
<td>65.66%</td>
<td>49.54%</td>
<td>69.16%</td>
<td>25.85%</td>
<td>86.33%</td>
</tr>
<tr>
<td>Caret</td>
<td>Random Forest Over 1:1</td>
<td>68.48%</td>
<td>43.48%</td>
<td>73.91%</td>
<td>26.56%</td>
<td>85.77%</td>
</tr>
<tr>
<td>Caret</td>
<td>XGBoost Boruta selected variables with Both 1:1</td>
<td>63.67%</td>
<td>52.29%</td>
<td>66.14%</td>
<td>25.10%</td>
<td>86.47%</td>
</tr>
<tr>
<td>Tidy Models</td>
<td>XGBoost Orginal</td>
<td>80.91%</td>
<td>0.58%</td>
<td>97.21%</td>
<td>31.15%</td>
<td>82.63%</td>
</tr>
<tr>
<td>Tidy Models</td>
<td>Random Forest Original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------</td>
<td>------------------------</td>
<td>-----------------------</td>
<td>----------------</td>
<td>------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>82.22%</td>
<td>1.68%</td>
<td>99.70%</td>
<td>55%</td>
<td>82.37%</td>
<td></td>
</tr>
<tr>
<td>Tidy Models</td>
<td>XGBoost Under 3:1</td>
<td>77.61%</td>
<td>17.13%</td>
<td>90.73%</td>
<td>29%</td>
<td>83.46%</td>
</tr>
<tr>
<td>Tidy Models</td>
<td>Random Forest Under 3:1</td>
<td>81.03%</td>
<td>8.05%</td>
<td>96.87%</td>
<td>35.83%</td>
<td>82.90%</td>
</tr>
<tr>
<td>Tidy Models</td>
<td>XGBoost Under 1:1</td>
<td>58.09%</td>
<td>59.99%</td>
<td>57.68%</td>
<td>23.53%</td>
<td>86.92%</td>
</tr>
<tr>
<td>Tidy Models</td>
<td>Random Forest Under 1:1</td>
<td>60.97%</td>
<td>61.01%</td>
<td>60.97%</td>
<td>25.33%</td>
<td>87.81%</td>
</tr>
<tr>
<td>Tidy Models</td>
<td>XGBoost Both 1:1</td>
<td>64.59%</td>
<td>50.46%</td>
<td>67.66%</td>
<td>25.294</td>
<td>86.29%</td>
</tr>
<tr>
<td>Tidy Models</td>
<td>Random Forest Both 1:1</td>
<td>67.51%</td>
<td>46.28%</td>
<td>72.12%</td>
<td>26.48%</td>
<td>86.08%</td>
</tr>
<tr>
<td>Tidy Models</td>
<td>XGBoost Over 1:1</td>
<td>66.76%</td>
<td>45.46%</td>
<td>71.39%</td>
<td>25.64%</td>
<td>85.78%</td>
</tr>
<tr>
<td>Tidy Models</td>
<td>Random Forest Over 1:1</td>
<td>69.93%</td>
<td>39.09%</td>
<td>76.62%</td>
<td>26.62%</td>
<td>85.29%</td>
</tr>
<tr>
<td>Tidy Models</td>
<td>XGBoost Boruta selected variables with Both 1:1</td>
<td>67.11%</td>
<td>46.79%</td>
<td>71.52%</td>
<td>26.28%</td>
<td>86.10%</td>
</tr>
</tbody>
</table>
### Table 12 10-fold Cross validation for Caret models predicting fatal overdose

<table>
<thead>
<tr>
<th>Model type</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive Predictive Value</th>
<th>Negative Predictive Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>71.80%</td>
<td>67.83%</td>
<td>75.80%</td>
<td>73.82%</td>
<td>70.08%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>74.75%</td>
<td>71.62%</td>
<td>77.90%</td>
<td>76.52%</td>
<td>73.18%</td>
</tr>
<tr>
<td>AdaBag</td>
<td>60.34%</td>
<td>64.70%</td>
<td>55.95%</td>
<td>59.63%</td>
<td>61.17%</td>
</tr>
<tr>
<td>SVM</td>
<td>59.35%</td>
<td>55.86%</td>
<td>62.86%</td>
<td>60.21%</td>
<td>58.61%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>60.60%</td>
<td>57.53%</td>
<td>63.69%</td>
<td>61.45%</td>
<td>59.85%</td>
</tr>
<tr>
<td>GBM</td>
<td>61.01%</td>
<td>58.61%</td>
<td>63.43%</td>
<td>61.72%</td>
<td>60.37%</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>58.99%</td>
<td>56.15%</td>
<td>61.86%</td>
<td>59.69%</td>
<td>58.37%</td>
</tr>
<tr>
<td>KNN</td>
<td>80.39%</td>
<td>71.05%</td>
<td>89.78%</td>
<td>87.49%</td>
<td>75.51%</td>
</tr>
<tr>
<td>Deep Learning</td>
<td>60.37%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### Table 13 10-fold Cross validation for Tidy models predicting fatal overdose

<table>
<thead>
<tr>
<th>Tidy Verse Models CV</th>
<th>Accuracy</th>
<th>Area Under the Receiver Operating Characteristic (AUROC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>75.40%</td>
<td>83.90%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>79.90%</td>
<td>88.30%</td>
</tr>
<tr>
<td>Model type</td>
<td>Accuracy</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>--------------------</td>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>60.60%</td>
<td>63.60%</td>
</tr>
<tr>
<td>KNN</td>
<td>76.60%</td>
<td>75.70%</td>
</tr>
<tr>
<td>XGBoost</td>
<td>63.54%</td>
<td>52.14%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>67.50%</td>
<td>45.46%</td>
</tr>
<tr>
<td>AdaBag</td>
<td>63.57%</td>
<td>55.41%</td>
</tr>
<tr>
<td>SVM</td>
<td>58.89%</td>
<td>57.95%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>59.89%</td>
<td>62.28%</td>
</tr>
<tr>
<td>GBM</td>
<td>60.69%</td>
<td>61.31%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>56.97%</td>
<td>60.91%</td>
</tr>
<tr>
<td>KNN</td>
<td>59.13%</td>
<td>49.75%</td>
</tr>
<tr>
<td>Deep Learning</td>
<td>71.19%</td>
<td>25.23%</td>
</tr>
</tbody>
</table>
Table 15 Test results from Tidy models for fatal overdose

<table>
<thead>
<tr>
<th>Tidy Verse Models</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive Predictive Value</th>
<th>Negative Predictive Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>64.59%</td>
<td>50.46%</td>
<td>67.66%</td>
<td>25.29%</td>
<td>86.29%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>67.26%</td>
<td>45.87%</td>
<td>71.91%</td>
<td>26.16%</td>
<td>85.96%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>59.89%</td>
<td>62.28%</td>
<td>59.37%</td>
<td>24.96%</td>
<td>87.88%</td>
</tr>
<tr>
<td>KNN</td>
<td>70.22%</td>
<td>23.90%</td>
<td>80.27%</td>
<td>20.82%</td>
<td>82.94%</td>
</tr>
</tbody>
</table>

Figure 19 Deep learning training and validation epochs for predicting fatal overdose
Figure 20 Area Under the Receiver Operating Characteristic (AUROC) for models predicting overdose
Figure 21 Forest plot of AUC scores from models predicting fatal overdose

5.5 Discussion

5.5.1 Model Performance Overview

The results from evaluating the class balanced samples show that utilizing both under and oversampling which led to a balanced count of 12,876 controls and 12,800 cases yielded the best results. This dataset was utilized for validation and testing of all models.

The predictive metrics for fatal overdose indicate that the machine learning models tested in this study are struggling to achieve high levels of predictive accuracy. The balance between sensitivity and specificity is a key challenge, with models like XGBoost and Random Forest showing moderate accuracy but lower sensitivity, which means they are missing a significant proportion of true fatal overdose cases.

The Random Forest model, exhibits a relatively higher specificity at 72.28% and an accuracy of 67.50%, indicating a better performance in correctly identifying those who did not
have a fatal overdose. However, its sensitivity of 45.46% points to the model's limitations in correctly identifying actual cases of fatal overdose. The Positive Predictive Value (PPV) across all models is relatively low, which suggests that when the model predicts a fatal overdose, it is correct only about a quarter of the time. This could be critical in a clinical setting where false positives can lead to unnecessary interventions. The AdaBag model shows a slightly better balance with a sensitivity of 55.41% and a specificity of 65.38%, but the improvements are marginal and still reflect significant room for improvement. The SVM model, while showing the highest PPV at 88.17%, is severely lacking in specificity (63.34%), suggesting that while it is confident in its predictions, it often falsely identifies non-cases as cases. The Logistic Regression and GBM models show some promise with higher sensitivity rates, indicating they are better at detecting true cases of fatal overdose. However, their lower specificity and PPV mean that there is a higher rate of false alarms, which could limit their practical application. Naïve Bayes and KNN models exhibit lower performance across most metrics, with KNN showing a particularly low sensitivity, implying that it misses more than half of the true fatal overdose cases. Deep Learning models, on the other hand, demonstrate high specificity but at the cost of sensitivity, which is not ideal for a condition where missing a true case can have dire consequences. The Tidyverse models reflect similar trends, with the XGBoost model showing a relatively balanced performance but still not at a level that would be considered clinically reliable.

Overall, the class imbalance seems to be a significant factor affecting the performance of these models. The underrepresentation of fatal overdose cases in the dataset can lead to models that are biased towards predicting the majority class—non-cases—more accurately. In predictive healthcare settings, this can be problematic because the cost of false negatives (missing true
overdose cases) is much higher than false positives. The resampling helped slightly with this but not enough.

5.5.2 Data and Model Challenges

The results from this study yielded results that suggest that these machine learning models fatal overdose a bit better than chance when looking at the overdose population. When compared to results from Chapter 2, these models are not performing well.

In discussing the challenges encountered in modeling fatal overdose risk, it is imperative to recognize the multifaceted nature of overdose as a phenomenon. The low predictive scores observed may largely be attributed to the intricate web of factors that contribute to an individual's risk of overdose, a complexity that may not be adequately represented in the available datasets.

A significant limitation in the data, as outlined in chapter 1.4.1, is the omission of crucial variables that characterize the overdose population. The absence of information on ethnicity and specific data regarding First Nations individuals hampers the model's ability to identify potential key predictors of overdose risk and restricts the development of culturally tailored interventions. This oversight not only impedes the precision of our predictive models but also highlights a broader issue of data representation and inclusivity. It should be clear that systematic exclusion through colonial oppression is the risk, not being a First Nations person. Furthermore, the lack of data on fentanyl use, a key driver in the recent surge of overdose deaths, is particularly concerning given its potency and rapid overdose potential.

Additionally, the models lack crucial personal behavioral traits, which literature identifies as risk factors for fatal overdoses. These include instances of using drugs alone, access to
naloxone, and proximity to healthcare facilities, all of which are vital in assessing the risk of fatal outcomes due to the absence of immediate assistance in emergencies (332). The predictive models' effectiveness is significantly undermined by not capturing these factors, especially since initial datasets from coroner services had this level of detail, but it was not available for overdose survivors. This discrepancy introduces a gap in understanding the conditions and behaviors leading to general overdoses, which may differ from those resulting in fatalities, further limiting the model's predictive capabilities.

Furthermore, the models may not capture the dynamic nature of overdose risk, which can change over time due to various factors, including personal circumstances, policy changes, and shifts in the drug supply. The predictive ability of models based on static historical data can be severely limited when applied to the current, rapidly evolving landscape of substance use.

The study also contends with an insufficient sample size of fatal overdose cases. The data spans only from 2015 to 2019, and the necessary practice of class balancing introduces a further complication: the test samples may not mirror the training data in distribution, potentially skewing the model’s predictive performance. This issue is exacerbated by a disproportionate representation of data from individuals who survived overdoses, as compared to those who did not, with the latter primarily derived from coroner’s services. Such an imbalance can bias the models, skewing them toward the predictors of fatal outcomes and overlooking the possibly distinct profiles of general overdoses.

The methodologies employed for data collection pose significant challenges in this study. While the use of passive big data collection offers a broad dataset, it lacks the detailed granularity required for models to achieve high-fidelity predictions. Critical and subtle predictors
of overdose may go undetected without actively collected data points that can capture the nuances of individual cases.

In the face of the drug toxicity crisis's continuous evolution, adopting a longitudinal analysis could provide a more nuanced view of the changing risk factors over time. Such an approach would allow for an in-depth examination of individual patient trajectories, healthcare utilization patterns, and the influence of shifting social determinants on the risk of overdose. However, the implementation of longitudinal analysis hinges on the initiation of more sophisticated data collection processes that actively track these variables over extended periods. Only with the establishment of such rigorous data collection protocols can there be development of models that truly encapsulate the complexities of the drug toxicity crisis.

The lack of incentives for data collection further complicates this picture. Frontline healthcare workers, such as ambulance services, are not sufficiently motivated to collect the detailed data that could be pivotal in refining the predictive models. As a result, the dataset might lack insightful information that these professionals are uniquely positioned to gather. Lastly, the general nature of the medical service plan and hospital data fails to capture the level of detail needed to accurately model overdose risks, such as the circumstances of each overdose event or detailed patient histories.

In sum, these limitations underscore the need for an enhanced data collection strategy and a more nuanced approach to model development. Improving the quality and specificity of the data collected, and subsequently enriching the modeling techniques, is essential for progressing toward more reliable predictive tools in the fight against the overdose epidemic.
Modeling for overdose: utilizing machine learning in predicting general overdose

5.6 Objective

This chapter describes the utilization of the BC Provincial Overdose Cohort (BCODC) collected by the BC Centre for Disease control to create multiple machine learning models predicting upon general overdose.

5.7 Methods

In this study, 'controls' refer to individuals who are healthy controls, while 'cases' denote individuals who experienced fatal and non-fatal overdose. All data analysis was conducted using RStudio (version 1.3.1093). Details of the coding packages/libraries employed, as well as the raw code, are provided in Appendix D. The analysis incorporated data that detailed the history of diagnoses spanning the years 2015 to 2019.

Table 16 Variables included in predictive modelling for general overdose

<table>
<thead>
<tr>
<th>Variables</th>
<th>Missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOH Study ID</td>
<td>0</td>
</tr>
<tr>
<td>Opioid Agonist Treatment</td>
<td>3306</td>
</tr>
<tr>
<td>Rheumatoid Arthritis</td>
<td>5175</td>
</tr>
<tr>
<td>Parkinsonism</td>
<td>5175</td>
</tr>
<tr>
<td>Osteoarthritis</td>
<td>5175</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>5175</td>
</tr>
<tr>
<td>Multiple Sclerosis</td>
<td>5175</td>
</tr>
<tr>
<td>Mood Anxiety</td>
<td>5175</td>
</tr>
<tr>
<td>Ischemic Stroke</td>
<td>5175</td>
</tr>
<tr>
<td>Ischemic Heart Disease</td>
<td>5175</td>
</tr>
<tr>
<td>Hypertension</td>
<td>5175</td>
</tr>
<tr>
<td>Hospital Transient Ischemic Attack</td>
<td>5175</td>
</tr>
<tr>
<td>Hospital Stroke</td>
<td>5175</td>
</tr>
<tr>
<td>Condition</td>
<td>Code</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Heart Failure</td>
<td>5175</td>
</tr>
<tr>
<td>Hemorrhagic Stroke</td>
<td>5175</td>
</tr>
<tr>
<td>Epilepsy</td>
<td>5175</td>
</tr>
<tr>
<td>Diabetes</td>
<td>5175</td>
</tr>
<tr>
<td>Depression</td>
<td>5175</td>
</tr>
<tr>
<td>Chronic Obstructive Pulmonary Disease</td>
<td>5175</td>
</tr>
<tr>
<td>Chronic Kidney Disease</td>
<td>5175</td>
</tr>
<tr>
<td>Asthma</td>
<td>5175</td>
</tr>
<tr>
<td>Angina</td>
<td>5175</td>
</tr>
<tr>
<td>Acute Myocardial Infarction</td>
<td>5175</td>
</tr>
<tr>
<td>Alzheimer Dementia</td>
<td>5175</td>
</tr>
<tr>
<td>Tobacco Use</td>
<td>1735</td>
</tr>
<tr>
<td>Tissue Infection</td>
<td>1735</td>
</tr>
<tr>
<td>Substance Related Disorders</td>
<td>1735</td>
</tr>
<tr>
<td>Stimulant Use</td>
<td>1735</td>
</tr>
<tr>
<td>Sepsis</td>
<td>1735</td>
</tr>
<tr>
<td>Sedative Hypnotic Use</td>
<td>1735</td>
</tr>
<tr>
<td>Psychotic Disorders</td>
<td>1735</td>
</tr>
<tr>
<td>Polysubstance Use</td>
<td>1735</td>
</tr>
<tr>
<td>Personality Disorders</td>
<td>1735</td>
</tr>
<tr>
<td>Other Psychoactive Drug Use</td>
<td>1735</td>
</tr>
<tr>
<td>Osteomyelitis</td>
<td>1735</td>
</tr>
<tr>
<td>Opioid Use</td>
<td>1735</td>
</tr>
<tr>
<td>Neurotic Related Disorders</td>
<td>1735</td>
</tr>
<tr>
<td>Neurocognitive Disorders</td>
<td>1735</td>
</tr>
<tr>
<td>Multiple Mental Illness</td>
<td>1735</td>
</tr>
<tr>
<td>Mood Disorders</td>
<td>1735</td>
</tr>
<tr>
<td>Intellectual Disability</td>
<td>1735</td>
</tr>
<tr>
<td>Hallucinogen Use</td>
<td>1735</td>
</tr>
<tr>
<td>Endocarditis</td>
<td>1735</td>
</tr>
<tr>
<td>Early Onset Disorders</td>
<td>1735</td>
</tr>
<tr>
<td>Developmental Disorders</td>
<td>1735</td>
</tr>
<tr>
<td>Cocaine Use</td>
<td>1735</td>
</tr>
<tr>
<td>Cannabinoid Use</td>
<td>1735</td>
</tr>
<tr>
<td>Behavioural Psychological Disturbances</td>
<td>1735</td>
</tr>
<tr>
<td>Alcohol Use</td>
<td>1735</td>
</tr>
<tr>
<td>general Overdose Case</td>
<td>0</td>
</tr>
</tbody>
</table>
5.7.1 Data Preprocessing

5.7.1.1 Sample size

Initially, the control group consisted of 608,182 individuals. From this, 36,679 individuals identified as part of the overdose (OD) population were removed, resulting in 571,503 unique control cases. Further refinement involved the exclusion of cases with missing data (125,333 individuals), leading to the omission of 162,012 individuals in total. This process yielded a final count of 446,170 unique healthy control cases. The study's final sample comprised these 446,170 healthy controls and 36,679 overdose cases, totaling 483,849 individuals. This number reflects the study's composition of healthy controls and overdose cases.

5.7.1.2 Data splitting

The train test split for machine learning was done utilizing the Classification And REgression Training (Caret), 6.0-94 package (312). Data was partitioned into 70% training data and 30% testing data. After partitioning, the train set had 338,373 observations, with controls being 312702 and cases being 25671. The test set had 145,016 observations, with controls being 134008 and cases being 11008. All data preprocessing was done after the split, to adhere to the golden rule of machine learning (313,314).

5.7.1.3 Missing Data

The overdose population cases have 1735 missing rows for data from MSP and DAD derived variables. 5175 missing rows for data from CDR. 3306 missing rows for OAT from Pharmanet.
After the train-test split, the train data have 1202 missing rows for data from MSP and DAD derived variables. 3612 missing rows for data from CDR. 2282 missing rows for OAT from Pharmanet.

After the train-test split, the train data have 533 missing rows for data from MSP and DAD derived variables. 1563 missing rows for data from CDR. 1024 missing rows for OAT from Pharmanet.

The Multiple Imputation by Chained Equations (MICE) package, version 3.15.0, in RStudio was utilized for handling missing binary outcome data. The adopted methodology was Classification and Regression Trees (CART) from the MICE package, which has demonstrated substantial improvements in imputation over traditional parametric methods due to its recursive partitioning approach (316). The number of imputations performed was set at 5 which was the most a column has for percentage of missing data, a quantity supported by the literature as robust (317–320).

5.7.1.4 Resampling

The training set was balanced by randomly undersampling, resulting in an equal number of controls and overdose cases. In contrast, the test set was deliberately left unbalanced to emulate real-world data more accurately, thus trying to reflect the actual distribution of overdoses within the general population. The methodology incorporated an understanding that, while the prevalence of overdoses in the general population is approximately 0.7%, the test data's 10% prevalence is a more realistic representation than a 50% prevalence. This approach, with a closer-to-real-world overdose prevalence in the test data, is designed to enhance the practical
applicability and accuracy of the predictive modeling, ensuring it is more aligned with real-world settings and overdose distribution patterns.

5.7.1.5 Machine learning

The methodology employed in this section replicates the approaches outlined in section 5.3.2.5.

5.7.1.5.1 Hyper parameter tuning

The methodology employed in this section replicates the approaches outlined in section 5.3.2.5.

5.7.1.5.2 Evaluation metrics

The methodology employed in this section replicates the approaches outlined in section 5.3.2.5.

5.7.1.5.3 Confidence intervals

The methodology employed in this section replicates the approaches outlined in section 5.3.2.5.

5.8 Results

5.8.1 Resampling results

The implementation of under sampling balanced the training data to controls and cases being 25,671, totaling 51342.
### 5.8.2 Machine learning results

Table 17 10-fold cross validation results for caret models predicting overdose

<table>
<thead>
<tr>
<th>Tidyverse Caret Model type CV</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive Predictive Value</th>
<th>Negative Predictive Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>85.69%</td>
<td>89.19%</td>
<td>82.19%</td>
<td>83.36%</td>
<td>88.38%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>85.53%</td>
<td>90.03%</td>
<td>81.03%</td>
<td>82.60%</td>
<td>89.04%</td>
</tr>
<tr>
<td>AdaBag</td>
<td>85.58%</td>
<td>89.14%</td>
<td>82.02%</td>
<td>83.22%</td>
<td>88.30%</td>
</tr>
<tr>
<td>SVM</td>
<td>84.96%</td>
<td>87.13%</td>
<td>82.79%</td>
<td>83.51%</td>
<td>86.55%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>85.60%</td>
<td>89.01%</td>
<td>82.18%</td>
<td>83.32%</td>
<td>88.21%</td>
</tr>
<tr>
<td>GBM</td>
<td>85.96%</td>
<td>87.98%</td>
<td>83.01%</td>
<td>83.82%</td>
<td>87.35%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>70.67%</td>
<td>95.50%</td>
<td>45.83%</td>
<td>63.81%</td>
<td>91.07%</td>
</tr>
<tr>
<td>KNN</td>
<td>83.28%</td>
<td>87.55%</td>
<td>79.02%</td>
<td>80.66%</td>
<td>86.38%</td>
</tr>
<tr>
<td>Deep Learning</td>
<td>72.12%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
### Table 18 10-fold cross validation results for tidy models predicting overdose

<table>
<thead>
<tr>
<th>Tidy Verse Models CV</th>
<th>Accuracy</th>
<th>Area Under the Receiver Operating Characteristic (AUROC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>85.50%</td>
<td>91.10%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>85.90%</td>
<td>91.70%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>85.60%</td>
<td>91.40%</td>
</tr>
<tr>
<td>KNN</td>
<td>76.10%</td>
<td>83.60%</td>
</tr>
</tbody>
</table>

### Table 19 Test results for caret models predicting overdose

<table>
<thead>
<tr>
<th>Tidyverse Caret Model type</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive Predictive Value</th>
<th>Negative Predictive Value</th>
<th>Area Under the Receiver Operating Characteristic (AUROC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>88.52%</td>
<td>81.69%</td>
<td>89.08%</td>
<td>38.06%</td>
<td>98.34%</td>
<td>90.84%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>88.77%</td>
<td>81.59%</td>
<td>89.36%</td>
<td>38.64%</td>
<td>98.34%</td>
<td>91.12%</td>
</tr>
<tr>
<td>AdaBag</td>
<td>89.19%</td>
<td>80.43%</td>
<td>89.91%</td>
<td>39.56%</td>
<td>98.24%</td>
<td>86.86%</td>
</tr>
<tr>
<td>SVM</td>
<td>86.35%</td>
<td>81.90%</td>
<td>86.72%</td>
<td>33.62%</td>
<td>98.32%</td>
<td>88.05%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>88.15%</td>
<td>81.50%</td>
<td>88.70%</td>
<td>37.20%</td>
<td>98.32%</td>
<td>90.67%</td>
</tr>
<tr>
<td>GBM</td>
<td>88.21%</td>
<td>82.32%</td>
<td>88.69%</td>
<td>37.42%</td>
<td>98%</td>
<td>91%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>93.61%</td>
<td>20.19%</td>
<td>99.64%</td>
<td>82.08%</td>
<td>93.83%</td>
<td>88.43%</td>
</tr>
<tr>
<td>KNN</td>
<td>87.02%</td>
<td>78.34%</td>
<td>87.74%</td>
<td>34.42%</td>
<td>98.01%</td>
<td>86.56%</td>
</tr>
<tr>
<td>Deep Learning</td>
<td>89.30%</td>
<td>71.25%</td>
<td>90.81%</td>
<td>38.91%</td>
<td>97.47%</td>
<td>85.50%</td>
</tr>
</tbody>
</table>
Table 20 Test results for tidy models predicting overdose

<table>
<thead>
<tr>
<th>Tidy Verse Models</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive Predictive Value</th>
<th>Negative Predictive Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>88.33%</td>
<td>81.51%</td>
<td>88.89%</td>
<td>37.61%</td>
<td>98.32%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>88.73%</td>
<td>81.92%</td>
<td>89.29%</td>
<td>38.59%</td>
<td>98.36%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>88.15%</td>
<td>81.50%</td>
<td>88.70%</td>
<td>37.20%</td>
<td>98.32%</td>
</tr>
<tr>
<td>KNN</td>
<td>88.43%</td>
<td>60.65%</td>
<td>90.71%</td>
<td>34.90%</td>
<td>96.56%</td>
</tr>
</tbody>
</table>

Table 21 Test results for Caret models predicting overdose with variables selected by Boruta.

<table>
<thead>
<tr>
<th>Model type</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive Predictive Value</th>
<th>Negative Predictive Value</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost (variables from CDR)</td>
<td>64.73%</td>
<td>71.26%</td>
<td>64.20%</td>
<td>14.05%</td>
<td>96.45%</td>
<td>76.11%</td>
</tr>
<tr>
<td>XGBoost (variables from MSP+DAD)</td>
<td>86.29%</td>
<td>83.51%</td>
<td>86.52%</td>
<td>33.73%</td>
<td>98.46%</td>
<td>89.22%</td>
</tr>
</tbody>
</table>
Figure 22 Deep learning training and validation epochs for predicting overdose
Figure 23 Area Under the Receiver Operating Characteristic (AUROC) for models predicting overdose
5.9 Discussion

5.9.1 Model Performance Overview

Across the various models, a general trend emerges where ensemble methods, including Random Forest and XGBoost, demonstrate superior performance compared to simpler models such as Naïve Bayes and KNN, particularly in terms of accuracy and Area Under the Receiver Operating Characteristic (AUROC). This suggests that the complexity of the data related to overdose prediction benefits from more sophisticated algorithms that can capture non-linear relationships and interactions between predictors.

Accuracy is a general indicator of a model's overall performance, while AUROC reflects its ability to distinguish between the positive (overdose) and negative (no overdose) classes. The Random Forest model appears to achieve the highest accuracy (88.77%) and AUROC (91.12%) among the Caret models, indicating a strong predictive power. XGBoost also performs well, with
comparable accuracy and AUROC scores, suggesting its robustness in handling the prediction task.

Sensitivity (true positive rate) and specificity (true negative rate) are critical when the costs of false negatives and false positives are high. In the context of overdose prediction, a high sensitivity means effectively identifying individuals at risk, which is crucial for timely intervention. The Random Forest model achieves the highest sensitivity (90.03%), indicating its effectiveness in correctly identifying overdoses. However, models like Naïve Bayes, despite having a lower accuracy, show a remarkably high sensitivity (95.50%), suggesting they could be useful in scenarios where missing an actual overdose case would be particularly detrimental, even though they might result in more false alarms (lower specificity).

The Positive Predictive Value (PPV) and Negative Predictive Value (NPV) provide insights into the performance of the model from a clinical utility perspective. PPV indicates the probability that subjects with a positive screening test truly have the condition, whereas NPV indicates the probability that subjects with a negative screening test truly don't have the condition. The models generally show high NPV but lower PPV, which is expected in settings where the prevalence of the condition (overdose) is low. This implies that while the models are reliable in identifying individuals who are not at risk, there is less certainty in the predictions of those at risk, highlighting the need for additional confirmatory tests or assessments in positive cases. In practical terms, this could lead to a situation where many individuals are identified as at-risk who are not actually at-risk, which could have significant implications for treatment or intervention strategies. This scenario is often observed in medical testing and screening where the prevalence of the condition is low. Even with high sensitivity and specificity, the chance of a
positive result being a true positive can be low simply because overdoses are rare when compared to the general population.

Deep learning models in this study demonstrate significant potential, as evidenced by their high accuracy rate of 89.30% in test results. This high accuracy indicates that these models are effective at correctly predicting outcomes in the dataset they were tested on. Additionally, these models also exhibit a relatively AUC score of 85.5%.

5.9.2 Comparative Analysis

Comparisons between Caret and Tidyverse implementations reveal similar performance among the models, with minor differences in accuracy and AUROC. This suggests that the selection of a particular package or implementation may not critically impact the model's outcomes, enabling researchers and practitioners to choose tools based on factors like user-friendliness, interpretability, and compatibility with current workflows.

In comparison to models that predict fatal overdoses, models predicting general overdoses show notably better performance. This improvement is attributed to the use of a much larger dataset, which effectively compensates for the limited detail available in the variables used.
Figure 25 Raw AUC scores from systematic review Chapter 2
The synthesis of results from models from this research project with those from the systematic review underscores the potential effectiveness of machine learning models in predicting overdoses. Models like Random Forest and XGBoost, which show both high individual performance and strong aggregated effect sizes, are promising tools for practitioners and researchers in the field of addiction psychiatry. The predictive models for overdose in this project exhibit exceptional performance, achieving an average AUC of 88.78%. This surpasses the benchmarks established by prior models in the field, which average 83% AUC, as depicted in Figure 26. However, models predicting fatal overdoses tend to mirror the lower range of outcomes seen in the Chapter 2 meta-analysis, exhibiting an average AUC score of 62.51% as depicted in Figure 25.

Furthermore, the relatively wide confidence intervals observed in the meta-analysis for some models indicate that while certain modeling approaches have the potential to be highly effective, their performance may depend heavily on the specifics of their application. This highlights the importance of model tuning, feature selection, and the consideration of population-
specific factors in the development of predictive models. The comparison suggests a convergence of evidence on the utility of certain models for predicting overdose events, with ensemble methods and deep learning models standing out. It also emphasizes the need for careful consideration of model choice based on the context and objectives of the predictive task at hand. The insights from the systematic review and meta-analysis reinforce the findings from individual studies and provide a comprehensive picture of the current state of overdose prediction modeling.

5.10 Conclusion

In conclusion, the choice of a predictive model for overdose should be guided by the specific context and the costs associated with different types of prediction errors. While ensemble methods like Random Forest and XGBoost offer strong overall performance, the high sensitivity of models like Naïve Bayes could be harnessed in scenarios where identifying as many true cases as possible is paramount. The high NPV across models is reassuring, suggesting that individuals identified as low risk are indeed unlikely to experience an overdose. However, the relatively low PPV across models necessitates cautious interpretation of positive results. Future work could focus on improving the PPV, potentially through the integration of additional predictive features or the development of hybrid models that combine the strengths of different algorithms. The integration of machine learning models into clinical practice for overdose prediction must also consider the ethical implications and the need for transparent and interpretable models.
Chapter 6: Exploratory secondary analysis, inference, and survival analysis

6.1 Objective

This chapter describes the utilization of the BC Provincial Overdose Cohort (BCODC) collected by the BC Centre for Disease control to create multiple inference-based models looking at general and fatal overdose. In additional, an exploratory survival analysis was created. The primary objective is to delve into the variables identified as significant in the machine learning predictive models discussed in the previous chapter. By doing so, this chapter will uncover trends and relationships between these variables and the outcomes of interest, thereby demystifying some of the 'black box' aspects inherent to machine learning methodologies. This approach allows for a more nuanced examination of the BCODC, facilitating a deeper comprehension of the complex interplay between the myriad factors at play and the resultant health outcomes observed in the cohort.

The aim of this chapter was to investigate features that were associated to fatal and general overdose. The inferential models applied here are designed to complement the predictive accuracy of machine learning with the explanatory power of traditional statistical methods. This chapter serves as a bridge between data-driven prediction and the clinical and policy-oriented application of those predictions to address the overdose crisis in British Columbia.

6.2 Stratifying risk factors in fatal overdose population

6.2.1 Methods

The analysis incorporated data that detailed the history of diagnoses spanning the years 2015 to 2019.
The methods used in the analysis included a variety of advanced statistical and machine learning techniques. Logistic regression was employed for inferential modelling, where the log odds of the outcome were taken and converted into odds ratios using the formula $OR = e^{\beta}$, with $\beta$ as the coefficient.

XGBoost was used to assess feature relevance, applying SHAP values for interpretability (SHAPforxgboost, version 0.1.3, and xgboost, version 1.2.0.1) (333). The Boruta algorithm (Boruta, version 8.0.0) was utilized to determine variable importance (334). The analysis also involved examining variable importance by incorporating Recursive Partitioning and Regression Trees as well as filter-based variable importance using the caret package (version 6.0-94) (335,336). In addition, recursive feature elimination (RFE) was also implemented through the caret package (version 6.0-94). Top 10 variables of importance was compared between these methodologies.

Additionally, to explore the relationships between variables, a correlational plot was created using Pearson's correlation coefficient, facilitated by the 'corrplot' package (version 0.84).

### 6.2.1.1 Fatal overdose dataset

For the analysis, a methodology that utilized the entire datasets for inference-based modeling was employed, instead of using a train-test split. This approach was specifically applied to the overdose dataset, as outlined in section 5 of the study.

The fatal overdose dataset comprised of 30,253 general overdose cases classified as controls, and 6,426 fatal overdose cases categorized as the case group, resulting in a total sample size of 36,679.
6.2.1.2 Missing Data

The approach for handling missing data for inference modelling followed the procedures detailed in Chapters 5 Specifically, the number of multiple imputations implemented was 15.

6.2.2 Results

6.2.2.1 Variable importance

Figure 27 Variable Importance in Fatal Overdose Prediction Using the Boruta Algorithm
6.2.2.2 Logistic Regression

Figure 28 Logistic Regression Odds Ratio of variable when predicting fatal overdose.
The raw scores of variable importance can be seen in the raw code in appendix D.

XGBoost, which stands for eXtreme Gradient Boosting, is an algorithm that has been utilized in chapter 5 and 6 as a predictive model. In this context, SHAP (SHapley Additive exPlanations) values was utilized, a method derived from game theory to explain individual predictions of machine learning models. They provide insights into how each feature in the dataset contributes to the model's prediction for a specific instance. This is particularly valuable
in complex models like XGBoost, where variable contributions can be hard to interpret. SHAP values help in interpreting the model by quantifying the impact of each feature on the prediction, thus offering transparency and aiding in the understanding of predictive models.

Figure 29 XGBoost Shap Value results for Fatal Overdose Prediction
Figure 30 SHAP values between total OD number and Multiple mental illnesses

Figure 31 SHAP values between total OD number and Mood Disorders
6.2.2.3 Correlational plot

![Correlational plot](image)

Figure 32 Correlational plot for fatal overdose (overdose population)

6.2.3 Discussion

The plots presented in this study offer a detailed overview of the risk factors influencing mortality, especially within the realm of substance use disorders. By analyzing the data from these graphs, insights can be gained into how chronic health conditions, mental health issues,
substance use, infectious diseases, and treatment interventions interact. Understanding these relationships is vital for devising focused strategies aimed at enhancing patient outcomes.

6.2.3.1 **Chronic diseases**

The increasing trajectory of conditions such as Acute Myocardial Infarction, Chronic Kidney Disease, and Chronic Obstructive Pulmonary Disease signals an escalating threat to survival as these conditions progress. This worsening impact emphasizes the necessity for an integrated healthcare approach that manages both substance use and chronic physical illnesses. The upward trend in coefficients for disorders like neurocognitive and mood disorders suggests that their influence on mortality risk intensifies over time. This pattern underlines the progressive nature of these conditions and the critical need for early mental health interventions and continuous support. The significant immediate effect of acute hospitalization events such as strokes and TIAs on mortality risk points to the crucial role of addiction psychiatry in managing these events. They can precipitate substance use relapses and should be addressed promptly within the context of substance use disorder management. The rising trend of conditions like endocarditis, often linked with intravenous drug use, highlights the cumulative effects of substance use and recurrent infections. This underscores the importance of infection control, harm reduction, and healthcare interventions in substance use disorder populations.

6.2.3.2 **Substance use and mental illness**

The different trajectories observed for alcohol, cannabis, and hallucinogen use indicate varying impacts on mortality. Initial substance use might increase risk, but this may change over time due to factors like treatment adherence or tolerance development.
The presence and trajectory of disorders like behavioral psychological disturbances and developmental disorders suggest that these conditions may chronically affect mortality risk. This effect could be direct or indirect, such as through decreased access to care or increased risk behaviors. The variability in the impact of Opioid Agonist Therapy (OAT) across models indicates the need for further research to understand the effectiveness of treatment interventions. The protective trend associated with OAT in some plots confirms its value in improving survival outcomes.

6.2.3.3 Figure interpretations

Integrating insights from various model outputs, such as XGBoost, Boruta Algorithm, RPART, Feature Importance, and Recursive Feature Elimination (RFE), provides a multi-faceted understanding of mortality risk factors. OAT and opioid use are recurrent themes across models, suggesting their significant impact on mortality outcomes. You can see the red from figure 28, which shows how OAT and being diagnosed with substance related disorders is negatively associated with fatal overdose. Moreover, the role of polysubstance use and other psychoactive drug use as consistent variables across different models underscores the complexity of substance use patterns and their effects on mortality.

The analysis of these data points reveals a complex landscape of factors affecting the mortality of individuals with substance use disorders. The dynamic interplay of chronic health conditions, mental and neurocognitive disorders, acute medical events, and substance use-related complications form a multifactorial web that healthcare providers must navigate. The evidence points towards the importance of comprehensive, integrated care that is responsive to the
evolving needs of this population. Early intervention, targeted treatment for substance use and co-occurring conditions, and ongoing support for chronic health and mental health issues appear to be key in improving survival outcomes. Additionally, the insights gathered from these analyses could inform public health strategies and policy-making, aiming to reduce mortality within substance-using populations by addressing these multifaceted risk factors in a holistic manner.
6.3 Stratifying risk factors in the general overdose

6.3.1 Methods

The analysis incorporated data that detailed the history of diagnoses spanning the years 2015 to 2019.

The methods used in the analysis included a variety of advanced statistical and machine learning techniques. Logistic regression was employed for inferential modelling, where the log odds of the outcome were taken and converted into odds ratios using the formula $\text{OR} = e^{\beta}$, with $\beta$ as the coefficient.

XGBoost was used to assess feature relevance, applying SHAP values for interpretability (SHAPforxgboost, version 0.1.3, and xgboost, version 1.2.0.1). The Boruta algorithm (Boruta, version 8.0.0) was utilized to determine variable importance. The analysis also involved examining variable importance by incorporating Recursive Partitioning and Regression Trees as well as filter-based variable importance using the caret package (version 6.0-94). In addition, recursive feature elimination (RFE) and random forest feature importance, both also implemented through the caret package (version 6.0-94). Top 10 variables of importance was compared between these methodologies.

Additionally, to explore the relationships between variables, a correlational plot was created using Pearson's correlation coefficient, facilitated by the 'corrplot' package (version 0.84).

6.3.1.1 Sample

For the analysis, a comprehensive methodology was utilized, involving the use of the entire datasets for inference-based modeling instead of a traditional train-test split. This method
was specifically applied to the overdose dataset, as elaborated in Chapter 6. The overdose dataset consisted of 446,710 individuals identified as healthy controls and 36,679 cases of overdose, amounting to a total of 483,389 individuals.

6.3.1.2 Missing data

The approach for handling missing data for inference modelling followed the procedures

6.3.2 Results

6.3.2.1 Variable importance

![Variable Importance in Overdose Prediction Using the Boruta Algorithm](image)

Figure 33 Variable Importance in Overdose Prediction Using the Boruta Algorithm
6.3.2.2 Logistic Regression

Figure 34 Logistic Regression Odds Ratio of variable when predicting overdose.
<table>
<thead>
<tr>
<th>XGBoost SHAP Values</th>
<th>Boruta Algorithm</th>
<th>RPART</th>
<th>FVI</th>
<th>RFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opioid Use</td>
<td>OAT</td>
<td>Polysubstance Use</td>
<td>Substance related disorders</td>
<td>Substance related disorders</td>
</tr>
<tr>
<td>Substance related disorders</td>
<td>Rheumatoid Arthritis</td>
<td>Substance related disorders</td>
<td>Polysubstance Use</td>
<td>Polysubstance Use</td>
</tr>
<tr>
<td>Polysubstance Use</td>
<td>Parkinsonism</td>
<td>Opioid Use</td>
<td>Opioid Use</td>
<td>opioid use</td>
</tr>
<tr>
<td>OAT</td>
<td>Osteo Porosis</td>
<td>OAT</td>
<td>OAT</td>
<td>OAT</td>
</tr>
<tr>
<td>Hypertension</td>
<td>Multiple Sclerosis</td>
<td>Stimulant use</td>
<td>Hypertension</td>
<td>Stimulant use</td>
</tr>
<tr>
<td>Tissue infection</td>
<td>Mood Anxiety</td>
<td>Other Psychoactive Drug Use</td>
<td>Mood Disorders</td>
<td>Hypertension</td>
</tr>
<tr>
<td>Stimulant Use</td>
<td>Ischemic Stroke</td>
<td>Cocaine Use</td>
<td>Stimulant Use</td>
<td>Tissue infection</td>
</tr>
<tr>
<td>Tobacco Use</td>
<td>Hypertension</td>
<td>Hypertension</td>
<td>Cocaine Use</td>
<td>Neurotic related disorders</td>
</tr>
<tr>
<td>Cocaine Use</td>
<td>Ischemic Heart Disease</td>
<td>Alzheimer/Dementia</td>
<td>Neurotic related Disorders</td>
<td>Asthma</td>
</tr>
<tr>
<td>Alcohol Use</td>
<td>Hospital Stroke</td>
<td>N/A</td>
<td>Other Psychoactive Drug Use</td>
<td>Mood disorders</td>
</tr>
</tbody>
</table>

The raw scores of variable importance can be seen in the raw code in appendix D.
Figure 35 XGBoost SHAP Value results for Overdose Prediction
6.3.2.3 Correlational plot

![Correlational plot for overdose]

Figure 36 Correlational plot for overdose

6.3.3 Discussion

The results from the inference-based models for predicting fatal overdose, as illustrated in Table 12, provide a substantive overview of the complexity and interplay of risk factors associated with fatal overdose. These models, which utilize a range of inferential statistical
techniques, reveal a nuanced set of risk factors that are both diverse and, in some instances, overlapping.

### 6.3.3.1 Chronic diseases

The inclusion of various chronic and acute medical conditions—like Alzheimer's Dementia, Acute Myocardial Infarction, Angina, CKD, COPD, Depression, Diabetes, and Heart Failure—in the Boruta algorithm's top risk factors underscores the interaction between physical health and the risk of fatal overdose. This suggests that managing substance use disorders in patients with comorbid medical conditions requires a comprehensive approach that considers the entire spectrum of patient health.

Early onset disorders appearing as a risk factor in multiple models indicates the potential long-term impact of early-life mental health challenges on the risk of fatal overdose. This could imply the importance of early intervention in such disorders to reduce the risk trajectory over a person's lifetime.

### 6.3.3.2 Substance use and mental illness

Opioid Agonist Therapy (OAT) features prominently across various inferential models, underscoring its significance in the context of substance use treatment. Its identification as a top risk factor suggests a possible correlation with high-risk individuals who are undergoing treatment. This finding necessitates a deeper exploration into the characteristics of those receiving OAT to better understand and mitigate the associated risk factors that contribute to the likelihood of a fatal overdose.
Mental health disorders—particularly neurotic-related disorders, and substance-related disorders—are recurrently identified as significant risk factors. This indicates a strong association between mental health comorbidities and the risk of fatal overdose. The consistent appearance of these factors across different inferential methods emphasizes the critical need for integrated treatment programs that address the multifaceted nature of substance use and mental health.

The identification of polysubstance use and the involvement of other psychoactive drugs as principal risk factors in several models stress the compounded risk associated with the concurrent use of multiple substances. This points to the necessity of addressing complex substance use patterns in prevention strategies.

The recurrence of total overdose number as a risk factor highlights the predictive value of previous overdose events, which may signal an escalating risk pattern leading to a fatal outcome. This underlines the importance of interventions following an overdose, such as increased monitoring and support, to prevent future fatal incidents.

6.3.3.3 Figure interpretations

The diversity of risk factors identified by these inferential models, which include Filter Variable Importance (FVI), speaks to the robustness of a multimodal inferential approach. The variable importance as determined by filtering methods provides a prioritized list of predictors based on their individual association with the outcome, independent of model structure. This allows for the identification of risk factors without the constraints of model assumptions.

However, these models are not without limitations. The absence of certain variables in specific models (indicated by N/A) may reflect methodological differences in feature selection
and importance ranking. This variation underscores the importance of triangulating findings across various inferential approaches to form a more comprehensive understanding of risk factors.

While these inferential models have provided valuable insights into the potential risk factors for fatal overdose, their exploratory nature warrants a cautious interpretation. The associations identified are suggestive rather than conclusive and should be validated in prospective studies designed to establish causality.

In summary, these inferential models have elucidated a spectrum of risk factors that are crucial for understanding and preventing fatal overdoses. The convergence of factors across models reinforces the need for a holistic and integrative approach to treatment and prevention that encompasses both mental and physical health considerations. The insights gained here lay a foundational framework for future research aimed at developing targeted and effective intervention strategies to mitigate the risk of fatal overdose.
6.4 Survival Analysis

6.4.1 Methods

6.4.1.1 Sample

The fatal overdose dataset was utilized for survival analysis, consisting of 30,253 general overdose cases (controls) and 6,426 fatal overdose cases (case group), totaling 36,679 cases. To incorporate a survival time component, the days survived were calculated from the onset of the opioid overdose crisis, designated as January 1st, 2015, until the time of death for each case. This approach allowed for a temporal analysis of survival within the context of the overdose crisis.

6.4.1.2 Packages used

Survival analysis was only done upon the dataset containing fatal and general overdose population. The survival package was utilized (survival, 3.3-5). Kaplan-Meier Estimator, Cox Proportional Hazards Model, Aalen's additive regression model and Random Forest were utilized to create survival curves for the overdose population.
6.4.2 Results

Figure 37 Individual patient survival curves from RF model

Figure 38 Survival curves of Cox, KM and RF models
Figure 39 Aalen's additive regression model for survival curves
Figure 40 Legend for Aalen's additive regression model for survival curves

Figure 41 Significant Aalen's additive regression model for total overdose number
6.4.3 Discussion

The application of Aalen's additive regression models has yielded insightful contributions to our understanding of the risk factors associated with mortality within the context of addiction psychiatry. This analysis reveals the temporal dynamics of how certain covariates impact survival probabilities.

The positive trajectories observed in figure 40 of the green plots for cocaine use, depression, and heart failure underscore a progressive escalation in the risk of mortality over time. This increment highlights the urgent need for continuous monitoring and intervention strategies to mitigate these risks. The growing coefficients suggest that the impact of these factors on mortality may intensify as time advances, prompting an exploration into the potential mechanisms that might underpin these associations.

In contrast, the downward trends in figure 40 of the blue and pink plots for OAT, opioid use, stimulant use, and substance-related disorders indicate a cumulative protective effect on survival. This is particularly salient in the context of OAT, where the negative coefficients may reflect the efficacy of such treatments in reducing mortality risk among patients with opioid dependence. The results align with existing literature that posits OAT as a cornerstone in the management of opioid dependence, potentially improving survival rates by reducing the risk of overdose and other adverse outcomes.

However, the interpretation of negative coefficients for opioid and stimulant use requires a nuanced understanding. The findings may at first seem paradoxical, given the known risks associated with substance use. Yet, these negative associations likely reflect the benefits of controlled, medically supervised interventions. This distinction underscores the importance of differentiating between illicit use and therapeutic use within clinical settings. The negative trend
associated with substance-related disorders may initially appear counterintuitive. Still, it could be indicative of successful engagement with treatment services or harm reduction strategies. It suggests that diagnosis and consequent treatment could inversely be associated with mortality, emphasizing the potential life-saving impact of early and sustained treatment interventions.

These findings must be contextualized within the complex landscape of addiction and its treatment. The covariates in this study, while statistically significant, represent just a snapshot of the multifaceted interactions between biological, psychological, and social factors that contribute to the prognosis of individuals with substance use disorders. Future research should continue to explore these dynamics over longer durations and in diverse populations to generalize these findings more robustly. Moreover, this dissertation highlights the utility of Aalen's additive model in addressing the nuances of time-varying effects in clinical research. By leveraging this modeling approach, the research offers a more granular perspective on how the risk factors for mortality evolve, providing valuable insights for clinicians and policymakers dedicated to reducing the burden of substance-related harm.

6.4.4 Limitations

This survival analysis presents several limitations that must be acknowledged to appropriately contextualize the findings. A significant limitation is the temporal framing of the study; the time component used in the analysis was not the onset of a specific disorder but could have been the point at which data became available for the opioid cohort. In theory, if individuals overdosed after 2015/01/01, then the onset of survival would be after their first overdose. This introduces a potential bias as the time of entry into the study may not accurately reflect the natural history or the chronological progression of the disorders being studied. Consequently,
survival times may not be representative of the true survival experience of individuals from the onset of their disorder.

Moreover, due to the retrospective nature of data collection, there could be a disparity in the timing of disorder onset and the initiation of treatment regimens, such as Opioid Agonist Therapy (OAT). This misalignment can obscure the precise estimation of treatment effects and the progression of comorbid conditions like heart failure, depression, and substance-related disorders. The impact of such time-related discrepancies can result in over or underestimation of the covariates' effects on the survival outcomes.

Additionally, the results of this study should be considered exploratory. The analysis serves as a preliminary investigation into the complex interplay between various factors and survival within an opioid cohort. The exploratory nature of this study implies that while the associations identified are suggestive, they are not definitive and should be interpreted with caution.

Future research should aim to utilize a longitudinal design where the time component corresponds with the onset of disorders to mitigate this bias. Such a design would allow for a more accurate assessment of the temporal sequence of events and the potential causal relationships between the covariates and survival outcomes. It would also be prudent for subsequent studies to employ a prospective approach, ensuring a more controlled and systematic collection of data to better establish temporal relationships and causality.
Chapter 7: Conclusions

Summary of Dissertation

The dissertation begins by contextualizing British Columbia's drug toxicity crisis within a historical framework, highlighting the systemic dissolution of institutional support and the rise of synthetic opioids like fentanyl, which exacerbated the crisis. It critiques the healthcare response, noting a missed opportunity for a comprehensive strategy proven effective in places like Switzerland and underscores the urgency for new strategies amidst rising fatalities. The introduction posits eHealth solutions, particularly the Risk Assessment and Management Platform (RAMP), as potential aids in leveraging data for personalized medicine to combat the opioid epidemic. A systematic review in Chapter 2 assesses machine learning models for opioid-related outcomes, revealing their promising efficacy, especially in predicting opioid use disorder (OUD) and overdose (OD). Chapter 3 evaluates clinical decision support systems (CDSS) in addictions, noting a shift towards data-driven models and their varied effectiveness in clinical settings. Chapter 4 details the exploratory data analysis and preparation for a pilot model using BC's overdose data, emphasizing the challenge of data imbalance and the strategic pivot required for comprehensive dataset construction. Chapter 5 discusses the moderate success of machine learning models in predicting fatal overdoses, hindered by data limitations, particularly the lack of detailed demographic and substance use information. Random forest performed the best with an area under curve metric of 63.52%. Chapter 5 extends the analysis to predicting general overdoses, showing improved model performance due to a larger dataset, yet still facing challenges in precision and the call for more sophisticated modeling techniques and enhanced data collection. Random forest performed the best with a area under curve metric of 91.12%. Chapter 6 conducts an in-depth analysis to uncover trends between variables significant in
predictive models and overdose outcomes, utilizing the BC Provincial Overdose Cohort. It employs logistic regression as well as other feature selection methodology to highlight key risk factors like Opioid Agonist Therapy and chronic diseases. Survival analysis reveals the impact of these covariates over time, emphasizing the need for integrated care and effective treatment interventions to mitigate overdose risk and improve survival outcomes. Together, these chapters demonstrate the complex yet promising role of machine learning in addressing the nuanced challenges of the opioid crisis, advocating for integrated clinical support systems and the importance of data quality and ethical considerations in predictive modeling.

**Summary of Introduction**

The introduction of this dissertation situates the drug toxicity crisis in BC within a historical and systemic framework. The dissolution of Riverview Hospital's institutional support led to a congregation of the mentally ill in the Downtown Eastside, where treatment availability transformed the area into a hub for the homeless and those seeking help (1,2). This period coincided with the rise of prescription opioid usage, prominently oxycontin, which eventually gave way to the more dangerous era of synthetic opioids, including fentanyl (6).

The response from healthcare authorities, focused primarily on harm reduction through safe supply, has not effectively implemented a comprehensive four-pillared approach that includes prevention, treatment, and enforcement along with harm reduction (337). Such an approach has been successfully adopted in other countries, notably Switzerland, suggesting a missed opportunity in BC (337). The consequences of this strategic shortfall are tragically quantified in recent fatality statistics. In 2022, the region experienced 2,377 deaths due to opioids. Alarmingly, by October 2023, the death count had reached 2,039, signifying a
deepening of the crisis, as depicted in Figure 4 of the dissertation (19). These figures call for a re-evaluation of current strategies and underscore the urgency of addressing this complex health emergency (19).

With the crisis continuing unabated and fatalities on the rise, the introduction shifts focus to the utilization of data amassed since 2015 by the BC Provincial Health Authorities after the declaration of a public health emergency in 2016 (16). Data collected into the BC Provincial Overdose Cohort and managed by the BC Centre for Disease Control, presents a significant opportunity for building eHealth solutions (43). Such solutions could play a pivotal role in intervention and expanding healthcare capabilities.

Within this context, Ehealth is a viable solution to help patients gain access to treatment (50). the Risk Assessment and Management Platform (RAMP) is highlighted as a promising application of machine learning. Machine learning could potentially harness this data to create risk assessment tools that integrate with current interventions, moving towards the goal of personalized medicine. By tailoring treatment and preventive strategies to the individual, there is potential to significantly improve the healthcare system's ability to respond effectively to the ongoing opioid epidemic.

**Summary of Chapter 2: Systematic review and meta-analysis on machine learning**

In this chapter, a systematic review and meta-analysis was done to investigate the effectiveness of various machine learning models in predicting opioid-related outcomes. The research encapsulated in Figure 11 evaluates the Area Under the Curve (AUC) of cohort studies through standardized mean differences, with a focus on outcomes like opioid use disorder
(OUD), opioid overdose (OD), and the risk of opioid use. The meta-analytic results point to a notable effect size averaging 1.28 (95% CI [1.14, 1.42]) across 28 studies involving 100 models. Despite the high heterogeneity observed in the study outcomes, the analysis revealed no evidence of publication bias according to Egger's regression. The subgroup analysis particularly highlighted that model predicting OUD and OD had more pronounced effect sizes than those estimating the risk of opioid use, suggesting their superior performance.

Figure 12 further categorizes the effect sizes based on the type of machine learning model employed. The overall effect size for these models was 1.27 (95% CI [1.12, 1.42]), derived from 28 studies encompassing 138 models, including logistic regression, decision trees/random forests, deep learning, and boosting algorithms. Similar to the findings in Figure 11, a high degree of heterogeneity was present, but Egger's regression confirmed an absence of publication bias. Among the model types, deep learning stood out with the highest effect size, followed by logistic regression and decision trees/random forests, with boosting algorithms also showing significant performance.

The chapter concludes with a comprehensive comparison of all models and outcomes based on their AUC performance, with detailed forest plots provided in Appendix A. This thorough analysis underscores the robust potential of machine learning in tackling the complex issue of opioid-related health outcomes, with specific model types offering particularly promising results.
Summary of Chapter 3: Systematic review and meta-analysis on clinical decision support systems

Chapter 3 of the dissertation presents a systematic review and meta-analysis of clinical decision support systems (CDSS) in the field of addictions and concurrent disorders. In this review, 69 studies were carefully examined, with the majority employing knowledge-based systems, while a growing number from 2017 onwards started to incorporate non-knowledge-based models. The review details the application of CDSS across various mental health conditions related to addictions and its concurrent disorders. The CDSS functions spanned from diagnosis and monitoring to medication management and follow-up care.

Until the year 2017, the development and use of Clinical Decision Support Systems (CDSS) in the field of addictions and concurrent disorders were focused solely on knowledge-based models. However, beginning in 2017 and continuing through 2022, there has been a noticeable emergence and inclusion of non-knowledge based, data driven models alongside the traditional ones. The chapter also provides an insightful analysis of the effectiveness of CDSS in Randomized Controlled Trials (RCTs), revealing an average efficacy effect size of -0.11, which, despite high heterogeneity, stands robust against publication bias. Satisfaction and acceptance outcomes related to CDSS usage showed a more substantial effect size of -0.50 and indicated potential publication bias.

Overall, Chapter 3 sheds light on the nuanced roles and outcomes of CDSS in mental health and addiction treatment, emphasizing their effectiveness in clinical settings while also highlighting areas for improvement in patient satisfaction and acceptance. The chapter
underscores the importance of CDSS in enhancing patient care and the potential for these systems to evolve and better serve the needs of those with addictions and concurrent disorders.

Summary of Chapter 4: EDA and pilot modeling

This chapter focuses on the critical task of data preparation and exploratory analysis within the context of BC's drug toxicity crisis. This chapter is pivotal as it transitions from the foundational understanding and evidence supporting machine learning's applicability, as established in previous chapters, to the practical application of these methodologies. Utilizing the comprehensive data collected by the BC Centre for Disease Control, the chapter outlines the process of transforming the BC Provincial Overdose Cohort (BCODC) data into a robust dataset for predicting general and fatal overdoses.

Access to an extensive array of healthcare datasets was granted through Population Data BC's secure research environment, ensuring data privacy and integrity. These datasets included a wealth of information ranging from medical service billings to mortality records, which are crucial for understanding the multifaceted nature of the drug toxicity crisis.

The chapter details the exploratory data analysis and the intricate process of data wrangling, utilizing an array of packages within the tidyverse suite in R (307). This process was essential in constructing an overdose case dataset, identifying individuals with overdose episodes and addressing the challenges of data imbalance and inconsistency initially encountered.

A significant portion of the chapter is dedicated to the development of a pilot model, which relied heavily on the Coroner's Service dataset. However, the need for a comprehensive dataset that equally represented general overdose cases led to a strategic pivot. The chapter
describes the selection and organization of five main datasets, each providing unique and valuable insights into the healthcare experiences of individuals who have overdosed.

Lastly, the chapter discusses the derivation of variables, employing methodologies to address the data inconsistencies and to construct a balanced and comprehensive dataset. This dataset formed the basis for the machine learning models developed to predict overdose outcomes, marking a critical step forward in the research project.

Summary of Chapter 5: predicting fatal overdose

Chapter 5 of the dissertation presents a detailed analysis of machine learning models in predicting fatal overdoses. The study involved a large sample of 36,679 cases, split into training and testing sets, with a focus on balancing the representation of general and fatal overdose cases. Despite using advanced algorithms like XGBoost, Random Forest, and others, the models generally achieved only moderate accuracy, with a tendency to miss a significant proportion of true fatal overdose cases (312).

The use of multiple imputation methods to address missing data, particularly with the MICE package, ensures that the data's integrity is maintained while accommodating real-world imperfections (315). The resampling approach, crucial in managing the class imbalance inherent in overdose data, involved random under sampling to achieve a more representative training dataset.

Key challenges identified encompass the lack of essential demographic data, especially concerning high-risk groups such as First Nations individuals affected by colonialism and colonial policies. Additionally, there is a notable absence of detailed information on substances,
particularly fentanyl, which plays a significant role in overdose fatalities. The study also encountered obstacles in data collection, predominantly relying on passively collected big data that did not provide the detailed granularity required for nuanced modeling.

The models' limitations in capturing the dynamic nature of overdose risk, along with the insufficient sample size of fatal cases and challenges in data collection methodology, were significant factors affecting their performance. The study suggests a need for enhanced data collection strategies and more sophisticated modeling techniques, emphasizing the importance of longitudinal analysis and active data collection to develop more reliable predictive tools for addressing the overdose epidemic.

Summary of Chapter 5: Predicting general overdose

Chapter 6 of the dissertation expands upon the methodologies applied in Chapter 5 but with a larger and more complex dataset to predict general overdose cases. The study analyzed a vast sample comprising 483,389 individuals, including 446,710 healthy controls and 36,679 overdose cases. This significant sample size, coupled with a careful train-test split and handling of missing data, sets a robust foundation for the machine learning analysis. Methods for dealing with missing values and class imbalance were the same as chapter 5.

The application of various machine learning models, including ensemble methods like Random Forest and XGBoost, as well as traditional algorithms like Naïve Bayes and KNN, provided a broad spectrum of insights into overdose prediction (312). The ensemble methods, known for their ability to handle complex datasets, demonstrated superior performance in terms
of accuracy and AUROC, suggesting their effectiveness in capturing the intricate patterns associated with overdose risks.

A key finding from this chapter is the generally better performance of models in predicting general overdoses compared to fatal overdoses, as seen in Chapter 5. This improvement can be attributed to the larger dataset available for general overdose prediction, which compensates for the lack of granularity in the variables. This indicates that dataset size and composition play a crucial role in the effectiveness of predictive models.

The chapter also offers a comparative analysis with findings from Chapter 2's systematic review. This comparison reinforces the potential of machine learning in overdose prediction, with ensemble methods and deep learning models showing promising results. The potential of these models, if integrated into a clinical decision support system as demonstrated in Chapter 3, could significantly impact clinical outcomes. Future research may explore the efficacy of these tools through randomized controlled trials to further validate their clinical utility.

In conclusion, Chapter 6 underscores the complexities and challenges in overdose prediction while demonstrating the efficacy of machine learning approaches. It highlights the importance of context-specific model selection, balancing the trade-offs between different types of errors, and the ethical considerations in deploying these models in clinical settings. The chapter calls for future work to focus on enhancing model precision, especially in improving PPV, and on exploring hybrid models that combine the strengths of various predictive algorithms.
Summary of Chapter 6: Exploratory analysis and survival analysis

Chapter 6 of the dissertation delves into exploratory secondary analysis, inference, and survival analysis, leveraging the BC Provincial Overdose Cohort collected by the BC Centre for Disease Control. It aims to explore variables significant in previous machine learning models, examining the complex interactions between various factors and overdose outcomes. This chapter transitions from predictive modeling to a more nuanced understanding of these relationships, enhancing the interpretability of machine learning results.

The analysis stratifies risk factors for fatal and general overdoses, employing logistic regression, XGBoost, the Boruta algorithm, and other statistical methods to identify key variables and their impact. Significant factors like opioid agonist therapy (OAT), substance-related disorders, and chronic diseases such as heart failure and depression are highlighted for their roles in influencing overdose risks. This multifaceted approach reveals the intricate dynamics between chronic health conditions, substance use, and mental health issues.

Survival analysis further investigates the temporal aspect of overdose risk, using Kaplan-Meier Estimators, Cox Proportional Hazards, and Aalen's additive models among others. This offers insights into how specific variables evolve over time in their impact on mortality, underscoring the importance of comprehensive treatment approaches. Notably, Aalen's model provides a nuanced view of how factors like cocaine use and OAT contribute to survival probabilities, emphasizing the progressive nature of risk associated with chronic and mental health conditions.
However, the chapter acknowledges limitations such as potential biases from the retrospective data collection and the exploratory nature of the analysis. It suggests a need for future research with a longitudinal design to better establish causality and the temporal progression of disorders.

In summary, Chapter 6 bridges the gap between data-driven predictions and their clinical applications, offering critical insights into the overdose crisis. It underscores the complexity of the issue, highlighting the need for integrated care strategies that address the interplay of substance use, mental health, and chronic diseases. This comprehensive analysis provides a solid foundation for future efforts to mitigate overdose risks through targeted interventions and policy-making.

7.1 Significance

7.1.1 Clinical Relevance

This project was designed to address the four different aspects of the continuum of care in addiction psychiatry [22]. This involves the clinical pillars of promotion, prevention, treatment, and recovery. The model allows for the promotion and knowledge translation of fatal and general overdose risk factors. It focuses on prevention efforts towards identified risk factors, which can help to reduce the risk of developing substance use disorders and other high-risk comorbidities among drug users. The project also can provide personalized treatment based on an individual’s characteristics and risk factor constellation. For example, it can link the identified risk behaviors to the appropriate treatment trajectories, such as the need for psychotherapy. Randomized controlled trials have supported behavioral and cognitive behavioral therapies in
their efficacy in addressing substance use disorders, which contribute to solving the drug toxicity crisis [23]. Lastly, this tool enables for the monitoring of risk behaviour that can assist in a long-term clinical trajectory. These aspects can be further enhanced on a web-based platform.

There are many benefits towards the utilization of a clinical virtual solution. For example, online tools are easily accessible, and are a low-barrier solution for connecting individuals with experts [3]. Virtual clinics have much shorter wait times in comparison to in-person clinics [3]. Moreover, these clinics facilitate the care of individuals with little-to-no access to care in rural areas, especially considering that travel times to receive treatment are a barrier [3]. This allows for a scalable web-based tool to expand its reach to marginalized populations without the need for additional human resources. Online platforms as a clinical solution allow individuals to safely access mental health resources without being stigmatized. Online health tools bolster the ability to monitor individuals by promoting their ability to log symptoms and behavioral changes. Lastly, such tools can facilitate peer mentorship and engagement among like-minded individuals. Recent studies show that a highly engaged group was uniquely associated with improving targeted goals and long-term life satisfaction [24]. This risk assessment tool allows individuals to compare their risk behaviors with their peers within the same population in order to promote healthier behaviors.

ML algorithms establish a strong analytical and conceptual platform for multiple data types and sources seeking virtual solutions [3]. Combining these robust tools alongside the existing mental health infrastructure (i.e., treatment trajectories and health care professionals) can create personalized medicine, which will significantly address the global mental health
burden of the drug toxicity crisis. Following the gold standard for disease and diagnosis in DSM-V and psychiatry, this model aims transform descriptive healthcare data at the individual level into behavioral patterns that can be monitored and analyzed in an effort to prevent risk for fatal and general overdose [25]. The key to substance use recovery is promoting an individualized treatment that includes comprehensive assessment and monitoring of symptoms and behavioral changes [25, 26].

There also exists the problem of non-modifiable risk factors. For example, there is the potential for variables selected as important by the ML models, such as ethnicity, age and gender, that cannot be changed. With the utilization of other methodologies, such as Bayesian Inference models, these risk factors can correlate upstream to prevent individuals already at risk. For example, a clinician could educate individuals about specific ethnic populations at risk for specific comorbidities. This allows for transparency so that the individual can be placed in a better position to deal with risks related to opioid overdose.

This tool attempts to harness the power of both collected healthcare data and expert opinion/literature to help build capacity at the individual, and health care levels. To be more precise, with the results of a ML algorithm, individuals will now have the opportunity to manage their risk behavior. With the development of a Clinical Decision Support System, individuals will be able to identify their risk and protective factors and share them with health care workers to help them to create an individualized treatment plan specific to the needs of the individual. The individual will finally be able address modifiable behaviors to decrease their risk of fatal overdose. In conclusion, this tool can play an important role by encouraging physicians to adopt a new gold standard of integrating personalized feedback research into clinical practice [27].
7.1.2 Research Relevance

In the existing research, different machine-learning models were created to address four other applications in mental health. These four applications include detection and diagnosis, prognosis treatment and support, public health applications, research, and clinical administration [28]. This project has focused on a combination of public health application domains, prognosis treatment, and support. The application of this project will also focus on research and clinical administration.

In literature, there is a demand for greater translational research to test the efficacy of ML models in psychiatry. The primary investigator’s preliminary results of two systematic reviews support this hypothesis. Results suggest that within 1966 studies identified in the space of ML and psychiatry, there were 46 ML algorithms created in the literature. However, when looking at how these algorithms translated in clinical studies, a second systematic review reported that within 6645 studies, 77 decision support systems were identified. Still, only 26 of them were ML-based. The rest of the 55 were rooted in knowledge-based expert opinions and literature. As a result, this project contributes to the additional efficacy studies of ML models with the end user, namely physicians and vulnerable populations in a clinical setting. To date, according to the primary investigator’s knowledge, this is the only project that utilizes data from the BC Provincial Overdose Cohort to create a CDSS with methods in ML in BC.

There is also the possibility of creating different models to target individual risk factors for diverse at-risk populations, such as youth. The literature supports that the current healthcare system’s response has been insufficient and needs substantial changes to address high-risk substance use and overdoses among youth [29]. In addition, there is also the ability to utilize the
same data set to predict individual comorbidity, such as suicide risk. Literature also supports comorbid threats of suicide among the opioid-using population [30]. Lastly, the combination of quantitative clinical data with qualitative data from neuroscience is a possibility. Although integrating multi-model clinical data can pose a significant challenge, the current literature suggests that it can significantly improve the performance of phenotyping and predictive algorithms [31].

In this section the student must demonstrate his/her mastery of the field and (for doctoral candidates) his/her contribution to knowledge in the broader discipline.

7.2 Limitations

In the limitations section of this study, it is crucial to address the inherent challenges and constraints associated with employing machine learning (ML) within the realm of healthcare data, particularly in the context of eHealth tools.

A fundamental principle to consider is 'garbage in, garbage out'. The effectiveness and accuracy of the predictions made by ML models are heavily reliant on the quality of the input data. Issues such as missing data, inaccuracies, biases, or data that is not representative of the target population can lead to erroneous model outputs. In the sensitive context of healthcare, such inaccuracies could significantly impact patient care and outcomes. Additionally, the limited availability of certain types of healthcare data, like long-term patient outcomes or detailed lifestyle information, can restrict the depth and accuracy of the inferences that can be drawn from these models.
The interpretability of ML models poses another significant challenge. Many advanced algorithms, particularly in deep learning, operate as 'black boxes', offering little insight into their decision-making processes. In healthcare, where understanding the reasoning behind diagnostic and therapeutic decisions is crucial, this lack of transparency can impede the trust and adoption of these models by healthcare professionals.

Another constraint is the generalizability of the models. ML models are frequently trained on specific datasets that may not be reflective of the broader population. Consequently, the applicability and relevance of the models' inferences may be limited to the specific context of the training data and may not extend accurately to different demographic or clinical settings. Furthermore, the risk of overfitting presents a substantial challenge. Machine learning models that are overly tailored to the training data may fail to generalize to new, unseen datasets. This limitation is exacerbated by the dynamic nature of healthcare data, where patient health trajectories and policy changes can swiftly render a previously effective model obsolete.

Furthermore, ML models, especially those used for predictive analytics, are not inherently designed for causal inference. They are adept at identifying correlations among variables, but these correlations do not necessarily imply a cause-and-effect relationship. Without robust experimental or longitudinal data to establish causal links, any conclusions drawn about the impacts of specific interventions or treatments based on ML models need to be approached with caution. Interpreting the significance of features identified by ML models is also complex. The importance attributed to a variable within a model does not automatically translate to clinical significance. Additionally, the interaction effects and non-linear relationships that complex models can capture often resist simple interpretation.
Computer science, while offering powerful tools for data analysis and model construction, brings its own set of limitations. Algorithmic bias, if unaddressed, can perpetuate disparities in healthcare delivery. Security and privacy concerns are paramount when dealing with sensitive health data, and the threat of data breaches remains a constant concern. Additionally, the integration and interoperability of data across diverse healthcare systems and formats present significant challenges to the cohesive analysis of health information. The data used in healthcare itself often suffers from issues of incompleteness and inconsistency. Missing values and non-standardized data entry practices can compromise the integrity of the datasets, upon which the machine learning models are built. Selective labeling of data, often a by-product of subjective clinical diagnoses, can introduce another layer of variability and potential error.

While this project was not significantly impacted by missing data, the methodological choice to oversample the fatally overdosed population introduced a bias. This overrepresentation of general overdose cases likely influenced the suboptimal performance of the predictive models when estimating fatal overdose outcomes.

When considering eHealth tools specifically, user engagement and adoption are critical for their success. Without consistent and correct usage by healthcare providers and patients, the effectiveness of these tools is compromised. Clinical integration of eHealth tools into existing workflows is crucial and often challenging to achieve. Scalability and maintenance are necessary to ensure the long-term viability and relevance of these tools, which can be resource intensive. Lastly, the digital divide cannot be overlooked. eHealth tools presuppose a level of access to technology and digital literacy that may not be universally available, potentially leading to disparities in access to these innovative healthcare solutions.
In addressing these limitations, this dissertation calls for a multi-faceted approach involving improved data collection and processing, enhanced model transparency, and more inclusive healthcare technology practices.

7.3 Future directions

Machine learning models designed to predict both fatal and general overdose outcomes will be retrieved from a secure research environment, ensuring data privacy and integrity. Once extracted, these models will be integrated into an Application Programming Interface (API). This API will serve as a conduit through which the Risk Assessment and Management Platform (RAMP) can access the predictive capabilities of the models. The RAMP, with its robust framework, is ideally positioned to leverage the predictive power of these models to identify individuals at heightened risk of overdose, facilitating timely interventions.

Furthermore, the application of inference-based models and the analysis of survival curves are instrumental in elucidating the multifactorial associations that exist between various determinants and the likelihood of fatal and general overdose incidents. By uncovering the significance and interaction of these risk factors, healthcare providers and policymakers can gain a deeper understanding of the dynamics at play, guiding the development of targeted prevention and treatment strategies.

The scope of these machine learning models extends beyond the realm of overdose prediction. The methodologies and insights garnered from this research hold the potential to be repurposed for forecasting other complex health outcomes and critical events, such as the occurrence of intricate disorders or the risk of suicide. Additional research aimed at customizing these models could pioneer the creation of comprehensive tools capable of predicting a broader
spectrum of adverse health events. This could significantly enhance the preventative measures in public health and clinical settings, enabling a more proactive approach to managing and mitigating such occurrences.

By continuing to refine these models and expand their applications, there exists a profound opportunity to harness machine learning not just as a predictive mechanism, but as a catalyst for proactive healthcare, delivering interventions that are as nuanced and complex as the conditions they aim to preempt.

7.4 Concluding remarks

The concluding remarks of this dissertation bring to light a stark reality: the current approach to addressing the needs of vulnerable, homeless PWUDs in BC is failing, as evidenced by the escalating numbers of unregulated drug deaths. It becomes increasingly clear that while harm reduction strategies, such as safe supply initiatives, are vital, they are not sufficient in isolation. A paradigm shift is essential, where machine learning and eHealth solutions emerge as critical components of a multifaceted response to this crisis.

These technologies offer an unparalleled opportunity to revolutionize the healthcare landscape by enabling the development of sophisticated predictive tools. Such tools can not only identify individuals at high risk of overdose but also facilitate the delivery of personalized, timely interventions. When integrated with the other fundamental pillars of addiction management—prevention and treatment—machine learning and eHealth can create a more holistic, effective strategy.

Leveraging the power of data and artificial intelligence enables a departure from generic solutions, steering towards a tailored understanding of individual circumstances. This approach
advocates for personalized interventions, aligning with the specific needs of each individual.

This shift towards a data-informed, personalized care framework could be the key to turning the tide on the drug toxicity crisis. It is a call for a renewed focus on comprehensive care that balances the immediacy of harm reduction with the long-term goals of health promotion and addiction recovery. The path forward is complex, but the integration of advanced technology with compassionate care holds the promise of a more hopeful future for those affected by addiction.
Bibliography


13. Williams E, Published HS. A Look at Changes in Opioid Prescribing Patterns in Medicaid from 2016 to 2019 [Internet]. KFF. 2023 [cited 2023 Dec 29]. Available from:


19. BC Corners. https://app.powerbi.com/view?r=eyJrIjoiY2ZkZTgxODAtMmE5Mi00MzNjLTlkNDYtMjRhNjU4NzkiNGZmIiwiCI6IjZmZGI1MjAwLTNkMGQtNGE4YS1iMDM2LWQzNgI1ZTM1OWFkYyJ9. 2023 [cited 2023 Dec 22]. Unregulated Drug Deaths- Summary. Available from: https://app.powerbi.com/view?r=eyJrIjoiY2ZkZTgxODAtMmE5Mi00MzNjLTlkNDYtMjRhNjU4NzkiNGZmIiwiCI6IjZmZGI1MjAwLTNkMGQtNGE4YS1iMDM2LWQzNgI1ZTM1OWFkYyJ9


44. BC Provincial Overdose Cohort (2017) [Internet]. Population Data BC; Available from: https://www.popdata.bc.ca/data


66. The potential use of telemedicine to increase access to opioid agonist therapy in rural and remote communities | British Columbia Medical Journal [Internet]. [cited 2024 Jan 4]. Available from: https://bcmj.org/mds-be/potential-use-telemedicine-increase-access-opioid-agonist-therapy-rural-and-remote


68. Wong JSH. A pilot study on risk and protective factors for opioid overdose among people who use street fentanyl during the COVID-19 pandemic [Internet]. University of British Columbia; 2021 [cited 2024 Jan 3]. Available from: https://open.library.ubc.ca/soa/cIRcle/collections/ubtheses/24/items/1.0398200


98. The HEALing (Helping to End Addiction Long-term SM) Communities Study: Protocol for a cluster randomized trial at the community level to reduce opioid overdose deaths through implementation of an integrated set of evidence-based practices. Drug and alcohol dependence. 2020;217.


111. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. 1997;30(7).


215. Iorfino F, Cross SP, Davenport T, Carpenter JS, Scott E, Shiran S, et al. A Digital Platform Designed for Youth Mental Health Services to Deliver Personalized and


detection and connected intervention to coproduce better care in severe mental illness. Studies

248. Bronzino JD, Morelli RA, Goethe JW. OVERSEER: a prototype expert system for
monitoring drug treatment in the psychiatric clinic. IEEE Transactions on Biomedical

medication treatment for opioid use disorder (PtDA-MOUD): Rationale, methodology, and

therapeutic response to oral ketamine for chronic suicidal ideation: a Bayesian network for

251. Tewari A, Kallakuri S, Devarapalli S, Jha V, Patel A, Maulik PK. Process evaluation of
the systematic medical appraisal, referral and treatment (SMART) mental health project in

Prospective evaluation of a clinical decision support system in psychological therapy. Journal

Psychopharmacological Treatment in the RAISE-ETP Study: Outcomes of a Manual and
1;175(2):169–79.

254. Chu KC, Huang YS, Tseng CF, Huang HJ, Wang CH, Tai HY. Reliability and validity of
DS-ADHD: A decision support system on attention deficit hyperactivity disorders. Computer
Methods and Programs in Biomedicine. 2017 Mar 1;140:241–8.

Decision Making for Antidepressants in Primary Care: A Cluster Randomized Trial. JAMA
Intern Med. 2015 Nov;175(11):1761–70.

Screening in Primary Care: Use of an Electronic Screener to Assess Suicidality and Improve

257. Røst TB, Clausen C, Nytrø Ø, Koposov R, Leventhal B, Westbye OS, et al. Local, Early,
and Precise: Designing a Clinical Decision Support System for Child and Adolescent Mental
Health Services. Frontiers in Psychiatry [Internet]. 2020 [cited 2023 Jul 18];11. Available

individualized digital decision assist system for the diagnosis and management of mental and


339. Tu K, Mitiku T, Lee DS, Guo H, Tu JV. Validation of physician billing and hospitalization data to identify patients with ischemic heart disease using data from the


## Appendices

### Appendix A Chapter 2 Appendix

#### A.1 Extraction Table

<table>
<thead>
<tr>
<th>Author (Publication year)</th>
<th>Demographics</th>
<th>Populaton</th>
<th>Features introduced to the ML model</th>
<th>Features selected by the ML model</th>
<th>Machine learning models and algorithms</th>
<th>Measures of model performance</th>
<th>Prediction and Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nath et al. [1]</td>
<td>Substance use population, Race, Age, Gender, Country of residence, and level of education</td>
<td>Substance use population</td>
<td>Personality: Agreeableness, conscientiousness, extraversion, neuroticism, openness to experience, impulsiveness, sensation seeking demographic: race, age, gender, country of residence, and level of education</td>
<td>Personality: Agreeableness, conscientiousness, extraversion, neuroticism, and openness to experience</td>
<td>Supervised Classification: Deep learning: artificial neural networks (ANN)</td>
<td>Accuracy</td>
<td>Classifying a drug user and predicting last drug use</td>
</tr>
<tr>
<td>Rice et al. [2]</td>
<td>Youth population, Age, Sex, State of residence (USA)</td>
<td>Youth population</td>
<td>Demographics: Age, Sex, State of residence (USA): ICD codes Prescription drug use: Opioid and derivatives, Buprenorphine, Methadone, Oxymorphone, Fentanyl, Morphine, Hydromorphone, Tramadol, Non-Opioid drug use, Comorbidities, and resource use</td>
<td>Chronic illness: High chronic pain, Demographics: Male gender and younger age, Mental Health: Higher mental health diagnosis, Medication: High opioid use and high prescription medications, such as antipsychotics</td>
<td>Supervised Classification: logistic regression</td>
<td>Accuracy AUROC</td>
<td>Risk of opioid abuse</td>
</tr>
<tr>
<td>Alghamdi et al. [3]</td>
<td>Hospital and health care population, Gender, Race and Level of education</td>
<td>Hospital and health care population</td>
<td>Demographics: Gender, Race and Level of education: Substance use: Cannabis use, age of first cannabis use, cannabis use frequency, cannabis use type, cannabis measure, cannabis use duration</td>
<td>Substance use: cannabis-related attributes</td>
<td>Supervised Classification: Support vector machines, Polynomial kernels, Gaussian Processes, Deep learning: Neural Network</td>
<td>Accuracy AUROC Sensitivity Specificity</td>
<td>First-episode psychosis</td>
</tr>
<tr>
<td>Ciesielski et al. [4]</td>
<td>Substance use population, Age, Sex, Region of USA</td>
<td>Substance use population</td>
<td>Demographics: Age: ICD-9 codes: Non-opioid substance abuse: 304.1 to 304.9 and 305.2 – 305.9, excluding 305.5, Tobacco use disorder: 305.1, Nondependent alcohol abuse: 303.9 and 305.0, Mental illness 290 to 302 and 306 to 316 Geography: USA regions Pharmacy claims: Morphine dose</td>
<td>Substance use: high morphine equivalent, Demographics: Younger age, Geography: South, West, and Midwest region, Mental health: Higher mental Health</td>
<td>Supervised Classification: Multivariate logistic regression</td>
<td>AUROC</td>
<td>ICD-9 diagnosis of non-dependent opioid abuse (304.0x) or dependent opioid abuse (305.5x)</td>
</tr>
<tr>
<td>Author (Publication year)</td>
<td>Demographic Features (Location (by ZIP code), Race/ethnicity, Gender, Age)</td>
<td>Clinical data (ICD 9 codes: Opioid abuse/dependence, drug abuse/dependence, alcohol abuse/dependence, mental health disorder, hepatitis c)</td>
<td>Subclinical data (Substance use: current smoker, Mental Health: mental disorder diagnoses)</td>
<td>Demographics: Age, area of residence, and risk factor method of payment, method of hospital admission</td>
<td>Demographics: Age, gender, ID number</td>
<td>Features introduced to the ML model</td>
<td>Features selected by the ML model</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------</td>
<td>----------------------------------------------------------------</td>
<td>-----------------------------</td>
<td>------------------------------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>Gutfraind et al. [5]</td>
<td>Substance use population, Location (by ZIP code), Race/ethnicity, Gender, Age</td>
<td>Injection drug use: Elapsed years of injection drug use, enrollment in any HR program, HCV infection state, Daily drug injections, sharing needles</td>
<td>Substance use: substance use severity</td>
<td></td>
<td></td>
<td>Substance use: substance use population, location</td>
<td>Substance use: substance use severity</td>
</tr>
<tr>
<td>White et al. [7]</td>
<td>Hospital and health care population, Substance use population, Age, Sex</td>
<td>Prescription drug use: Number of pharmacies, early prescription opioid refills, Number of prescriptions, early prescription opioid refills, and dose escalation</td>
<td>Substance use: opioid abuse diagnosis and positive smoking status</td>
<td></td>
<td></td>
<td>Substance use: opioid abuse diagnosis and positive smoking status</td>
<td>Demographics: Male gender and Ages between 18-34</td>
</tr>
<tr>
<td>Hylan et al. [8]</td>
<td>Hospital and health care population, Age, Gender, Ethnicity</td>
<td>Clinical data: clinician-documented problem opioid use Demographics: Age, Gender, Ethnicity</td>
<td>Substance use: opioid abuse diagnosis and positive smoking status</td>
<td></td>
<td></td>
<td>Substance use: opioid abuse diagnosis and positive smoking status</td>
<td>Demographics: Ages between 8-44</td>
</tr>
<tr>
<td>Choi et al. [9]</td>
<td>Hospital and health care population, Substance use population, age</td>
<td>Clinical data: Toxic substance (ICD-10 codes T65.0-T65.9)</td>
<td>Substance use: toxic substance, the intent of intoxication, severity, and risk factor</td>
<td></td>
<td></td>
<td>Substance use: toxic substance, the intent of intoxication, severity, and risk factor</td>
<td>Demographics: Youth</td>
</tr>
<tr>
<td>Che et al. [10]</td>
<td>Hospital and health care population, Substance use population, Age, Sex, Race, Mortality</td>
<td>Clinical data: ICD-9, ICD-10, HICD, CPT/HCPCS, RxNorm Code</td>
<td>Substance use: Mortality, Tobacco Use</td>
<td></td>
<td></td>
<td>Substance use: Mortality, Tobacco Use</td>
<td>Demographics: Sex, Age, Race, Mortality</td>
</tr>
<tr>
<td>Author (Publication year)</td>
<td>Demographic Populations</td>
<td>Features introduced to the ML model</td>
<td>Features selected by the ML model</td>
<td>Machine learning models and algorithms</td>
<td>Measures of model performance</td>
<td>Prediction and Outcomes</td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------------------------</td>
<td>-------------------------------------</td>
<td>----------------------------------</td>
<td>---------------------------------------</td>
<td>-----------------------------</td>
<td>------------------------</td>
<td></td>
</tr>
<tr>
<td>Olivia et al. [11]</td>
<td>Hospital and health care population, Substance use population, Sex, Age</td>
<td>Hospital and health care population, Sex, Age</td>
<td>Chronis Health: AIDS, blood loss anemia, Cardiac arrhythmia, chronic obstructive pulmonary disease, congestive heart failure, coagulopathy, deficiency anemia, diabetes, electrolyte disorder, hypertension, hypothyroidism, liver disease, lymphoma, metastatic cancer, obesity, osteoporosis, neurological disorder, paralysis, peptic ulcer, peripheral vascular disorder, a pulmonary circulation disorder, renal failure, rheumatoid arthritis, sleep apnea, solid tumor, valvular disease, Weight loss</td>
<td>Substance use: Drug dose</td>
<td>Supervised Classification: Multivariate mixed-effects logistic regression</td>
<td>AUROC Risk for overdose or suicide-related events</td>
<td></td>
</tr>
<tr>
<td>Zvi et al. [12]</td>
<td>Hospital and health care population, Substance use population, Sex, Age</td>
<td>Hospital and health care population, Substance use population, Sex, Age</td>
<td>436 predictor candidates Clinical data: ICD-9 ICD-10 for Opioid use disorder, RXCUI Codes for Opioid use disorder, ICD-9 ICD-10 for psychoactive substances use or dependence, abuse of nonpsychoactive substances, Alcohol use or dependence, Tobacco/ Nicotine use or dependence, Cannabis use or dependence, non mood psychotic disorders, mood (affective) disorders, Nonpsychotic mental disorders, Behavioral syndromes, Personality disorders, Impulse disorders, Alcohol-induced mental disorders, Drug-induced psychotic disorders, Migraine, Back pain, Fibromyalgia, Headache, Opioid adverse effects, Poisoning by opioids and narcotics, Other drug/ substance-related overdose, Diseases of the musculoskeletal system and connective tissue, Pain disorders, Postoperative pain, pain due to trauma, Temporomandibular Joint Disorders, Osteoarthritis, Kidney or gallbladder stones, Menstrual or genital pain, Scoliosis, Ankylosing spondylitis, Spinal arthropathy, Spondylitis, Gout, Hyperuricemia, Deposition Disease, Neuropathies, Gene deficiency anemia, diabetes, electrolyte disorder, hypertension, hypothyroidism, liver disease, lymphoma, metastatic cancer, obesity, osteoporosis, neurological disorder, paralysis, peptic ulcer, peripheral vascular disorder, a pulmonary circulation disorder, renal failure, rheumatoid arthritis, sleep apnea, solid tumor, valvular disease, Weight loss</td>
<td>Substance use: preoperative opioid duration, antidepressant use, tobacco use</td>
<td>Supervised Classification: Gradient boosting Deep learning: Word2vec</td>
<td>AUROC Negative predictive value Precision Sensitivity Specificity Detection of opioid use disorder</td>
<td></td>
</tr>
<tr>
<td>Zdilfer et al. [13]</td>
<td>Hospital and health care population, Age, Sex, Race, Marital status, Body mass index, US census region</td>
<td>Hospital and health care population, Sex, Age</td>
<td>Demographics, the Charlson Comorbidity Index (CCI) score, individual CCI comorbidities, other selected pain- and non-pain-related comorbidities, prescription medication information, including the opioid active ingredient, formulation, and MED, and select concomitant medications known to potentiate opioid effects, and health care utilization</td>
<td>Substance use: prescription opioid, concomitant prescribed benzodiazepines or antidepressants Chronic diseases: renal, liver, and pulmonary comorbidities and active traumatic injury Mental Health: Mental health disorders including opioid dependence</td>
<td>Supervised Classification: logistic regression</td>
<td>AUROC Other Measurements Risk prediction for Overdose</td>
<td></td>
</tr>
<tr>
<td>Author (Publication year)</td>
<td>Demographic Population</td>
<td>Features introduced to the ML model</td>
<td>Features selected by the ML model</td>
<td>Machine learning models and algorithms</td>
<td>Measures of model performance</td>
<td>Prediction and Outcomes</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------------------</td>
<td>-------------------------------------</td>
<td>----------------------------------</td>
<td>----------------------------------------</td>
<td>-----------------------------</td>
<td>------------------------</td>
<td></td>
</tr>
<tr>
<td>Zheng et al. [14]</td>
<td>Hospital and health care population, Age, Gender</td>
<td>Hospital and health care population</td>
<td>2186 features with p-value &lt; 0.05: Diagnostic codes ICD, Demographic features, medical prescription, Procedural codes, Emergency room visits, Social determinants</td>
<td>Diagnostic code (Substance): Substance abuse Serotonin Reuptake Inhibitor Acute alcoholic intoxication Alcohol-related disorders Demographics: Female gender Vehicle accidents Vehicle accidents at home</td>
<td>Supervised Classification: XGBoost, logistic regression Deep learning: Neural network AUROC Other Measurements</td>
<td>Predicting suicide attempt</td>
<td></td>
</tr>
<tr>
<td>Fitzgerald et al. [15]</td>
<td>Youth Population, Hospital and health care population, Age, Gender</td>
<td>Youth Population, Hospital and health care population</td>
<td>Individual Demographics: Gender (% Female) School Year (M/SD) Ethnic Minority (% Caucasian) Seen Mental Health Professional (% Yes), APSS-3 Psychotic Symptoms (M/SD) DASS-21 Depression (M/SD) DASS-21 Anxiety (M/SD) DASS-21 Stress (M/SD) CSI-15 Avoidance Coping (M/SD) CSI-15 Planning Coping (M/SD) CSI-15 Support Coping (M/SD) Anger (%Sometimes % Yes) Body Dissatisfaction (M/SD) Social determinants: BAS-7 Acting Out Behavior (M/SD) BMSLSS Satisfaction with Life (M/SD) LOT-R Optimism (M/SD) READ-Social Competence (M/SD) RSE Self-esteem (M/SD) Family Maternal Employment (% Employed) Paternal Employment (% Employed) Substance use: Alcohol, Cannabis, Tabacco Personality: Trouble with police Substance use: Alcohol, Cannabis, Tabacco</td>
<td>Supervised Classification: logistic regression with Elastic Net regularization</td>
<td>AUROC Other Measurements</td>
<td>Predicting alcohol, tobacco, and cannabis use</td>
<td></td>
</tr>
<tr>
<td>Author (Publication year)</td>
<td>Demographic Population</td>
<td>Features introduced to the ML model</td>
<td>Features selected by the ML model</td>
<td>Machine learning models and algorithms</td>
<td>Measures of model performance</td>
<td>Prediction and Outcomes</td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------------------------</td>
<td>------------------------------------</td>
<td>----------------------------------</td>
<td>---------------------------------------</td>
<td>----------------------------</td>
<td>------------------------</td>
<td></td>
</tr>
<tr>
<td>Lo-Ciganic et al. [16]</td>
<td>Substance use population, Age, Sex, Race/ethnicity, Disability, as the reason for Medicare eligibility, Receipt of low-income subsidy, and urbanicity of county of residence</td>
<td>Substance use population</td>
<td>Substance use: high-dose use, defined as higher than 120MME for 90 or more continuous days, four or more opioid prescribers and four or more pharmacies, concurrent opioid and benzodiazepine use for 30 or more cumulative days</td>
<td>Supervised Classification: random forest, gradient boosting machine, Multivariate logistic regression, most minor absolute shrinkage, and selection operator-type regression, Deep learning: Deep neural network</td>
<td>AUROC, Negative predictive value, Precision, Sensitivity, Specificity, Other Measurements</td>
<td>Risk prediction for overdose</td>
<td></td>
</tr>
<tr>
<td>Afshar et al. [17]</td>
<td>Hospital and health care population, Age, Sex, Gender</td>
<td>Hospital and health care population</td>
<td>Substance use: terms such as 'heroin,' 'opiates,' 'drug abuse,' and 'polysubstance abuse'</td>
<td>Supervised Deep learning: Convolutional neural network</td>
<td>AUROC, Negative predictive value, Precision, Sensitivity, Specificity, Other Measurements</td>
<td>Validating opioid misuse in hospital</td>
<td></td>
</tr>
<tr>
<td>Fulton et al. [18]</td>
<td>Hospital and health care population, Substance use population, Net Income</td>
<td>Hospital and health care population</td>
<td>Clinical: Staffed Beds, Surgeries, Affiliated Physicians</td>
<td>Supervised Classification: Linear regression, Lasso regression, Ridge regression, Elastic net</td>
<td>AUROC</td>
<td>Predicting hospital Opioid admissions</td>
<td></td>
</tr>
<tr>
<td>Cox et al. [19]</td>
<td>Substance use population, Age, Sex, Number of families, Household income</td>
<td>Substance use population</td>
<td>Clinical data: 3,000 variables related to SUDs and other psychiatric disorders, other health-related behaviors, and demographic variables with opioid cessation among EAs and AAs assessed in a cross-sectional study of opioid, cocaine, and/or alcohol dependence</td>
<td>Supervised Classification: Support vector machine, most minor absolute shrinkage and selection operator-type regression, random forest Deep learning: deep neural network</td>
<td>Other Measurements</td>
<td>Substance use</td>
<td></td>
</tr>
<tr>
<td>Chartash et al. [20]</td>
<td>Hospital and health care population</td>
<td>Hospital and health care</td>
<td>Clinical data: ICD and Billing codes</td>
<td>Supervised Classification: logistic regression</td>
<td>Negative predictive value, Precision, Other Measurements</td>
<td>Opioid Use Disorder</td>
<td></td>
</tr>
<tr>
<td>Author (Publication year)</td>
<td>Demographic Population</td>
<td>Features introduced to the ML model</td>
<td>Features selected by the ML model</td>
<td>Machine learning models and algorithms</td>
<td>Measures of model performance</td>
<td>Prediction and Outcomes</td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>------------------------</td>
<td>-------------------------------------</td>
<td>----------------------------------</td>
<td>--------------------------------------</td>
<td>-----------------------------</td>
<td>------------------------</td>
<td></td>
</tr>
<tr>
<td>Dong et al. [21]</td>
<td>Hospital and health care population, Age, Sex, Gender, Ethnicity</td>
<td>Hospital and health care population</td>
<td>Diagnosis (n = 457): First three digits of ICD-10 codes (ICD-19 codes were first converted to ICD-10 codes) Medications (n = 227): The total quantity of a medication a patient received Clinical events (n= 251): The highest, lowest and median values for each event if there are multiple values in one encounter Laboratory tests (n= 530): The numbers of high, low, and normal values and the total number for each test</td>
<td>Chronic pain, Dorsalgia, general pain, acute abdomen pain, joint disorder, issue disorder, hypertension, cystic fibrosis, type 2 diabetes, nausea, and vomiting Substance use: Opioid, Long term drug therapy, Nicotine dependence, Drug used in addictive disorders</td>
<td>Supervised Classification: Random forest, Decision tree, Logistic regression Deep learning: Long short-term memory, dense neural network</td>
<td>AUROC, F1 Precision Risk of opioid use disorder for patients taking opioid medications</td>
<td></td>
</tr>
<tr>
<td>Lo-Ciganic et al. [23]</td>
<td>Hospital and health care population, Age, Sex, Race, Type of Medicaid eligibility, and Duration of continuous enrollment</td>
<td>Hospital and health care population</td>
<td>Demographics: Age sex, race, type of Medicaid, duration of enrollment Clinical data: Allegheny County data Mental health: Health Status factors Substance use: Patterns of prescription opioid use</td>
<td>Demographics: Age, Race, Gender, Public benefit services Substance use: Opioid use disorder, Cumulative duration of gabapentinoid use Medical: Continuous Medicaid coverage, Medicaid eligibility, number of outpatient visits, Community services programs</td>
<td>Supervised Classification: Gradient Boosting Model</td>
<td>AUROC Predicting risk of opioid overdose among Medicaid beneficiaries</td>
<td></td>
</tr>
<tr>
<td>Author (Publication year)</td>
<td>Demographic information</td>
<td>Populaiton</td>
<td>Features introduced to the ML model</td>
<td>Features selected by the ML model</td>
<td>Machine learning models and algorithms</td>
<td>Measures of model performance</td>
<td>Prediction and Outcomes</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------------------</td>
<td>------------</td>
<td>-----------------------------------</td>
<td>---------------------------------</td>
<td>--------------------------------------</td>
<td>-----------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Dong et al. [24]</td>
<td>Hospital and health care population</td>
<td>Youth population</td>
<td>Substance use population, Age, Gender, Race</td>
<td>Substance use population n, Substance use population n</td>
<td>SPARCS Dataset: Diagnostic codes, Procedure codes (surgical, medical or diagnostic interventions received by patients), Demographic Information</td>
<td>Supervised Classification: Random Forest, Decision Tree, Logistic regression</td>
<td>Accuracy AUROC F1 Precision Recall</td>
</tr>
<tr>
<td>Choi et al. [25]</td>
<td>Youth population</td>
<td>Substance use population, Age, Gender, Ethnicity, Race, Language, Disability</td>
<td>Demographics: Age, Gender, Ethnicity, Race, Language, Disability</td>
<td>Substances use: Tobacco use, frequency and type</td>
<td>Demographics: Age, Gender, Ethnicity, Race, Language, Disability</td>
<td>Supervised Classification: Random Forest, LASSO</td>
<td>Accuracy Other measurements</td>
</tr>
<tr>
<td>Ahn et al. [26]</td>
<td>Substance use population, Age, Education, IQ, Family History, Sex</td>
<td>Substance use population n</td>
<td>Demographics: Age, Education, IQ, Family history, Sex Neurocognitive</td>
<td>Demographics: Low IQ, Low education</td>
<td>Supervised Classification: Elastic net</td>
<td>AUROC</td>
<td>Identifying markers for opiate and stimulant dependence</td>
</tr>
<tr>
<td>Afrizet et al. [27]</td>
<td>Canadian and Australian cohort Youth population</td>
<td>Substance use population, Sex, Age, Immigrant background</td>
<td>Demographics: Sex, age, immigrant background Mental health (Strengths and Difficulties Questionnaire (SDQ)): four neurocognitive and psychological domains of peer problems, conduct problems, hyperactivity, and emotional problems. Personality: anxiety sensitivity, negative thinking, impulsivity, and sensation seeking</td>
<td>Immigrant background protective factor</td>
<td>Supervised Classification: Support vector machine, random forests, Logistic regression, Lasso regression, Ridge regression, Elastic net Deep learning: Neural network</td>
<td>AUROC</td>
<td>Predicting adolescent alcohol use</td>
</tr>
<tr>
<td>Zhang- James et al. [28]</td>
<td>Swedish Youth population, age</td>
<td>Swedish Youth population n</td>
<td>Clinical data Demographics: Medical birth register: Sex, Age, Mother’s ID, etc. National school register: Merit score, secondary education, age at graduation, highest education level achieved, National Crime Register</td>
<td>Substance use: SUD Demographics: Nonviolent crimes, Violent crimes, Family income, NDEP Mental health: ADHD, Psychostimulant treatment</td>
<td>Supervised Classification: Random Forest, Deep learning: Recurrent neural network</td>
<td>AUROC</td>
<td>Predicting comorbid substance use disorders in ADHD in Youth</td>
</tr>
<tr>
<td>Author (Publication year)</td>
<td>Demographic Population</td>
<td>Features introduced to the ML model</td>
<td>Features selected by the ML model</td>
<td>Machine learning models and algorithms</td>
<td>Measures of model performance</td>
<td>Prediction and Outcomes</td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>------------------------</td>
<td>-------------------------------------</td>
<td>----------------------------------</td>
<td>----------------------------------------</td>
<td>-----------------------------</td>
<td>------------------------</td>
<td></td>
</tr>
<tr>
<td>Morel et al. [29]</td>
<td>United States Hospital and health care population, Age, Sex</td>
<td>Longitudinal Integration Database for Health Insurance and Labor Market Studies, Substance use: Prescribed Drug Register, National Patient Register</td>
<td>Clinical: Index hospital length of stay, discharge disposition (discharged to other institutional care versus home), Number of days and Number of times admitted to the in-patient hospitals in the previous 365 days (about 12 months), diagnosis category classified by the CCS and the DRG, index hospital stay expenditure, provider geographic location (state), Mental Health: Number of mental and medical comorbidities (CS), and age group</td>
<td>Supervised Boruta Algorithm Classification: XGBoost, GLMNet</td>
<td>AUROC</td>
<td>Predicting hospital readmission in mental/Substance use population</td>
<td></td>
</tr>
<tr>
<td>Ellis et al. [30]</td>
<td>United States Hospital and health care population, Substance use population, Age, Sex, Gender</td>
<td>Clinical Data: admission source, principal diagnosis based Clinical Classification Software (CCS), secondary diagnosis based Comorbidity Software (CS), diagnosis-related groups (DRG), history of previous M/SUD related hospitalization, length of index hospital stay and expenditure, discharge disposition (other institutions or home), geographic location of providers</td>
<td>Mental Health: major depression diagnosis and trazodone prescription (used to treat major depression)</td>
<td>Supervised Classification: Random Forest</td>
<td>AUROC</td>
<td>Predicting opioid dependence</td>
<td></td>
</tr>
<tr>
<td>Dong et al. [31]</td>
<td>United States Hospital and health care population, Age, Sex, Race, Marital status</td>
<td>Laboratory tests (n= 394): Hemoglobin, Red blood cell distribution, MME, Ultrasound, Chronic Illness: HIV diagnosis</td>
<td>Substance use: Opioids, Alcohol use, Smoking, Antipropulsives, Anesthetics, Other analgesics and antipyretics</td>
<td>Supervised Deep learning: Long Short term memory, Dense neural network Classification: Random forest, Decision tree, Attention, Logistic regression</td>
<td>AUROC F1</td>
<td>Predicting opioid overdose risk in patients with opioid prescriptions</td>
<td></td>
</tr>
<tr>
<td>Karhade et al. [32]</td>
<td>Hospital and health care population, age, sex, marital status</td>
<td>Demographics (age, sex, marital status, veteran status, race, ethnicity), disposition (inpatient, outpatient), neighborhood (zip-code) characteristics (median household income, median age, high school graduation or General Equivalency Diploma [GED] attainment, insurance status (Medicaid, Medicare, workers compensation, uninsured, private)</td>
<td>Substance use: benzodiazepine use, antidepressant use, gabapentin use</td>
<td>Supervised Classification: Random forest, stochastic gradient boosting, support vector machine, Elastic-net penalized logistic regression Deep learning: Neural Network</td>
<td>AUROC Other measurements</td>
<td>Predicting prolonged opioid prescriptions</td>
<td></td>
</tr>
<tr>
<td>Author (Publication year)</td>
<td>Demographic Population</td>
<td>Features introduced to the ML model</td>
<td>Features selected by the ML model</td>
<td>Machine learning models and algorithms</td>
<td>Measures of model performance</td>
<td>Prediction and Outcomes</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------------</td>
<td>-----------------------------------</td>
<td>----------------------------------</td>
<td>--------------------------------------</td>
<td>-----------------------------</td>
<td>------------------------</td>
<td></td>
</tr>
<tr>
<td>Glenz et al. [33]</td>
<td>Hospital and health care population, Substance use population, Age on index date, Gender, Race/ethnicity, and Medicaid coverage</td>
<td>Demographics: Age, Gender, Race/ethnicity, Medicaid coverage, Substance use: substance use disorders, LA/ER opioid formulations, Tobacco use</td>
<td>Mental health: mental health diagnoses, Substance use: substance use disorders, LA/ER opioid formulations, Tobacco use</td>
<td>Supervised Classification: Cox proportional hazards regression</td>
<td>Accuracy AUROC Sensitivity Specificity</td>
<td>Two-Year Risk of Opioid Overdose Among Patients Prescribed Chronic Opioid Therapy</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>, unemployment rate, population density [per square mile])</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>operative diagnosis (disk herniation, spondylolisthesis, spinal stenosis)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>procedural factors (fusion, approach, instrumented fusion, multilevel intervention)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>history of any previous spine surgery in the year before the index procedure, laboratory values (white blood cell count [×10^3] per microliter [μL]):</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>hemoglobin (grams per deciliter [g/dL]), platelet count (×10^3/μL), creatinine (mg/dL), preoperative medications (angiotensin-converting enzyme [ACE] inhibitor, angiotensin II receptor blocker [ARB], antidipressant, beta-2-agonists, beta-blockers, benzodiazepines, gabapentin, immunosuppressants, nonsteroidal anti-inflammatory drugs, antipsychotics)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>preoperative comorbidities (tobacco use, alcohol abuse, drug abuse, diabetes, renal failure, depression, psychoses, myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular accident, chronic obstructive pulmonary disease, arrhythmias, valvular disease, liver disease, and malignancy)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hospital and health care population, Substance use population, Age on index date, Gender, Race/ethnicity, and Medicaid coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Demographics: Age, Gender, Race/ethnicity, Medicaid coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Substance use: substance use disorders, LA/ER opioid formulations, Tobacco use</td>
<td>Mental health: mental health diagnoses, Substance use: substance use disorders, LA/ER opioid formulations, Tobacco use</td>
<td>Supervised Classification: Cox proportional hazards regression</td>
<td>Accuracy AUROC Sensitivity Specificity</td>
<td>Two-Year Risk of Opioid Overdose Among Patients Prescribed Chronic Opioid Therapy</td>
<td></td>
</tr>
<tr>
<td>Author (Publication year)</td>
<td>Demographic Population</td>
<td>Features introduced to the ML model</td>
<td>Features selected by the ML model</td>
<td>Machine learning models and algorithms</td>
<td>Measures of model performance</td>
<td>Prediction and Outcomes</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------------------</td>
<td>------------------------------------</td>
<td>----------------------------------</td>
<td>----------------------------------------</td>
<td>----------------------------</td>
<td>------------------------</td>
<td></td>
</tr>
<tr>
<td>Calcaterra et al. [34]</td>
<td>Hospital and health care population, Substance use population, Gender, Race, Age, Medicare, Medicaid</td>
<td>Clinical data: Milligrams of morphine per hospital, Demographics: Age, Insurance</td>
<td>Demographics: Age, Discount payment plan for insurance, Medicare, Commercial insurance</td>
<td>Supervised Classification: Stepwise Regression</td>
<td>Accuracy, AUROC, Sensitivity, Specificity</td>
<td>Predicting Future Chronic Opioid Use Among Hospitalized Patients</td>
<td></td>
</tr>
<tr>
<td>Gradus et al. [35]</td>
<td>Demark Hospital and health care population, Age, Marital status, Income</td>
<td>1339 variables spanning domains of suicide risk factor (<a href="https://jamanetwork.com/journals/jamapsychiatry/fullarticle/27530145">https://jamanetwork.com/journals/jamapsychiatry/fullarticle/27530145</a> note-Y13198064-1)</td>
<td>Chronic illness: Injury to the knee and lower leg, Infection of the skin and tissue.</td>
<td>Supervised Classification: Classification tree model (CART)</td>
<td>AUROC</td>
<td>Predicting Sex-Specific Suicide Risk</td>
<td></td>
</tr>
<tr>
<td>Meghan et al. [36]</td>
<td>Substance use population, Age, Gender, Race, Marital status, Employment, Service connection, Combat history</td>
<td>Chronic Illness: Number of Remissions, Number of pan diagnoses, Charlson index</td>
<td>Mental health: Cocaine abuse</td>
<td>Supervised Classification: Logistic Regression</td>
<td>Not Reported</td>
<td>Identifying predictors of Resolution of Aberrant Drug Behavior in Chronic Pain Patients Treated</td>
<td></td>
</tr>
<tr>
<td>Thornton et al. [37]</td>
<td>United States Hospital and health care population, Substance use population, age, sex, and US region (East, Midwest, South, or West)</td>
<td>Opioid regimen characteristics: Initial opioid prescriptions, Dose Pain Conditions: Chronic, Acute, Arthritis Physical and mental health conditions: Mental illness, Concomitant medications: Benzodiazepine use, Stimulant use, non-opioid analgesic, polypharmacy Demographics: Sex, Age, Region, Insurance plan type</td>
<td>Prescription drug use: Long-acting opioid, Tramadol, Opioids, Oxycodone, Hydrocodone, Benzodiazepine, High dose</td>
<td>Supervised Classification: Decision tree, Logistic regression</td>
<td>AUROC</td>
<td>Identifying predictors of Transitioning to Incident Chronic Opioid Therapy</td>
<td></td>
</tr>
<tr>
<td>Author</td>
<td>Demographic Population</td>
<td>Features introduced to the ML model</td>
<td>Features selected by the ML model</td>
<td>Machine learning models and algorithms</td>
<td>Measures of model performance</td>
<td>Prediction and Outcomes</td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>------------------------</td>
<td>--------------------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------</td>
<td>---------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Sharma et al.</td>
<td>Hospital and health care population, Age, Sex</td>
<td>Hospital and health care population, Clinical data: CNN CUI model had a total number of 5,721,449 trainable parameters, and the max-pooling network CUI model had 4,401,257 trainable parameters</td>
<td>Demographics: Victim of abuse, Substance use: Heroin, Cocaine, Methadone, Opiates, Albuterol, Dilaudid, Drug overdose, Oxycodone</td>
<td>Supervised Deep learning: Convolutional neural network, Deep averageing network, Max pooling network, Deep averageing + max pooling</td>
<td>AUROC</td>
<td>Identifying opioid misuse</td>
<td></td>
</tr>
<tr>
<td>Wadekar et al.</td>
<td>Substance use population, Age, Gender, Race</td>
<td>Substance use population, Opioid Dependence or Abuse in the past year Demographics: Male/Female, Age, Race, Income, Employment, Education, Approach by someone selling drugs, Heroin fairly or very easy to obtain, Any Disability, Any mental illness in the previous year, First use of alcohol before 18 years, eight years and non-users, Overall Health, Probation status in the past year, Perception of significant risk trying heroin once, Perception of significant risk using heroin weekly, Obesity: BMI</td>
<td>Demographics: approached by someone selling drugs, Age, Perception that drugs are easy to obtain, income, education, employment, race, Sub stance use: Early initiation of marijuana before 18, Early initiation of alcohol before 18, Chronic illness: Obesity, Disability</td>
<td>Supervised Classification: Logistic Regression, Decision Tree, Random Forest</td>
<td>Accuracy, AUROC, Specificity</td>
<td>Predicting adults at risk for OUD</td>
<td></td>
</tr>
<tr>
<td>Stichting et al.</td>
<td>Substance use population, Age, Sex, Race, Geographic region</td>
<td>Substance use population, Personality: Changes in Sleeping Pattern Crying, Loss of Interest in Sex, Changes in Appetite, Concentration Difficulty, Irritability, Self Dislike, Self Criticalness, Loss of Interest in Sex, Indecisiveness, Loss of Pleasure, Loss of Energy, Tiredness, Punishment Feelings, Guilty Feelings, Agitation, Past Failure, Worthlessness, Sadness, Pessimism</td>
<td>Personality: Crying, Suicidal Thoughts, Worthlessness, Pessimism, Loss Of Interest, Pessimism, Changes In Appetite, Indecisiveness, Agitation, Tiredness</td>
<td>Supervised Classification: Generalized additive model</td>
<td>AUROC</td>
<td>Predicting cocaine use frequency from depressive symptoms</td>
<td></td>
</tr>
<tr>
<td>Lee et al.</td>
<td>Hospital and health care population, Substance use population, Race, Sex, Education</td>
<td>Hospital and health care population, The 179 attributes were entered into the WEKA tool: one attribute was treatment-seeking status, and the other 178 attributes were used to train the ADT</td>
<td>Mental health: Mood Substance use</td>
<td>Supervised Classification: Random forest, Random tree, Logistic model, Deep learning: ADT</td>
<td>AUROC</td>
<td>Classifying Alcohol Use Disorder</td>
<td></td>
</tr>
<tr>
<td>Author (Publication year)</td>
<td>Demographic population</td>
<td>Population</td>
<td>Features introduced to the ML model</td>
<td>Demographics: Sex, Race/ethnicity, Class standing, Grade 8 or lower, Grades 9-12, High school diploma or above, Household income, Less than $20,000, $20,000 - $49,999, $50,000 - $74,999, $75,000 or more, Number of people living in household, Covered by Medicaid/CHIP, Overall health status, Exposure to primary prevention programming and efforts, Drug education in school, Drug program outside of school, Substance use program/counseling, STD/PREG prevention program, Violence prevention program, Problem solving group, Seen drug prevention, Talking to parents about danger of substance use, Youth experience, Youth activity participation, number of fights with parents, Parents’ help with homework, Parents’ limit going out with friends, Major depressive episode, Emotional treatment, Non-specialty mental health service, Tobacco use experience, Alcohol use experience, Marijuana use experience</td>
<td>Features selected by the ML model</td>
<td>Machine learning models and algorithms</td>
<td>Measures of model performance</td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------------------</td>
<td>------------</td>
<td>-------------------------------------</td>
<td>---------------------------------</td>
<td>-------------------------------</td>
<td>------------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Han et al. [42]</td>
<td>Youth population, Age, Education, IQ, Gender, Race</td>
<td>Youth population</td>
<td>Depression, Age, Education, IQ, Gender, Race, Class standing, Grade 8 or lower, Grades 9-12, High school diploma or above, Household income, Less than $20,000, $20,000 - $49,999, $50,000 - $74,999, $75,000 or more, Number of people living in household, Covered by Medicaid/CHIP, Overall health status, Exposure to primary prevention programming and efforts, Drug education in school, Drug program outside of school, Substance use program/counseling, STD/PREG prevention program, Violence prevention program, Problem solving group, Seen drug prevention, Talking to parents about danger of substance use, Youth experience, Youth activity participation, number of fights with parents, Parents’ help with homework, Parents’ limit going out with friends, Major depressive episode, Emotional treatment, Non-specialty mental health service, Tobacco use experience, Alcohol use experience, Marijuana use experience</td>
<td>Substance use: Marijuana use experience, Tobacco use experience, Alcohol use experience</td>
<td>Supervised Classification: distributed random forest, and gradient boosting machine</td>
<td>Accuracy</td>
<td>Predicting diagnosis of Opioid Use Disorder</td>
</tr>
<tr>
<td>Lo-Ciganic et al. [43]</td>
<td>United States Youth population, Substance use population, Sex, Race, Class, Overall health</td>
<td>United States Youth population, Substance use population</td>
<td>Patterns of prescription opioid use, Patterns of non-opioid prescription use ICD-10, Beneficiaries, Demographics, Health status factors, Opioid prescribing-level variables, Regional-level factors <a href="https://journals.plos.org/plosone/article?id=10.1371/journal.pone.02359319">https://journals.plos.org/plosone/article?id=10.1371/journal.pone.02359319</a></td>
<td>Geography: Reside in Nevada Prescription drug use: Schedule IV short-acting opioids Substance use: Elithalshuer drug abuse, Non-opioid drug disorders, alcohol use disorder Chronic illness: lower back pain, Fibromyalgia, Disabled, Arthritis, liver diseases Demographics: Race black</td>
<td>Supervised Classification: elastic net</td>
<td>AUROC</td>
<td>Other measures Predicting risk of incident opioid use disorder</td>
</tr>
<tr>
<td>Zedler et al. [44]</td>
<td>United States Substance use population, Age, Sex, Race, Disability, Medicaid, Medicare</td>
<td>United States Substance use population</td>
<td>Demographic factor, comorbidities of the Charlson Comorbidity Index, other selected pain- and nonspain-related health conditions, prescription medication information, including opioid active ingredient, formulation, and maximum prescribed daily morphine equivalent dose (MED), non-opioid medications that may impact serious adverse opioid effects (non-opioid analgesics, benzodiazepines, antidepressants, muscle relaxants, other sedatives, antipsychotics, and stimulants), metrics indicative of health care utilization</td>
<td>Substance use: Substance use disorder Mental health: Bipolar disorder/schizophrenia Clinical: Heart failure, Renal disease with renal impairment Prescription drugs: Morphine, Fentanyl, Methadone, Benzodiazepines, Antidepressants</td>
<td>Supervised Classification: Multiple Logistic Regression</td>
<td>AUROC</td>
<td>Validating screening risk index for overdose</td>
</tr>
<tr>
<td>Barak-Cohen et al. [45]</td>
<td>United States Hospital and Clinical data: ICD codes over 3000 Medication</td>
<td>United States Hospital and Clinical data</td>
<td>Mental health: Past suicide attempt, anxiety disorder</td>
<td>Supervised Classification: Naïve Bayes Classifier</td>
<td>AUC, Sensitivity, PPV, NPV</td>
<td>Suicide Risk Prediction Modeling</td>
<td></td>
</tr>
<tr>
<td>Author (Publication year)</td>
<td>Demographic Population</td>
<td>Features introduced to the ML model</td>
<td>Features selected by the ML model</td>
<td>Machine learning models and algorithms</td>
<td>Measures of model performance</td>
<td>Prediction and Outcomes</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------------------</td>
<td>--------------------------------------</td>
<td>----------------------------------</td>
<td>--------------------------------------</td>
<td>-----------------------------</td>
<td>-------------------------</td>
<td></td>
</tr>
<tr>
<td>Hospital and health care population Substance use population, Age, Sex, US Region</td>
<td>health care population Substance use population</td>
<td><a href="https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2763237">https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2763237</a></td>
<td>Substance use: Substance abuse Chronic illness: Arthritis excluding, Back pain, Vehicle accidents at home, Headache, PTSD Demographics: Female</td>
<td>Supervised Classification</td>
<td>AUROC</td>
<td>Predicting risk for Opioid Use Disorder</td>
<td></td>
</tr>
<tr>
<td>Walsh et al. [46]</td>
<td>Hospital and health care population Substance use population, Sex, Race</td>
<td>Hospital and health care population Clinical data: ICD codes over 5000 Diagnostic data grouped to Centers for Medicare &amp; Medicaid Services Hierarchical Condition Categories (HCC) (eg, schizophrenia-related ICD codes mapped to HCC 57) Medication data grouped to the Anatomic Therapeutic Classification, level IV (eg, citalopram N06AB04 [level V] maps to Selective Serotonin Reuptake Inhibitors N06AB [level IV]); Past health care utilization (counts of inpatient, emergency department, and ambulatory surgery visits over the preceding five years); Area Deprivation Indices33 by patient zip code</td>
<td>Demographic: Age Substance use: medical history of substance abuse Mental Illness: Mood disorder, anxiety disorder Chronic Illness: low back pain, renal impairment, painful neuropathy, and recent ER visit</td>
<td>Supervised Classification</td>
<td>AUROC</td>
<td>Predicting risk for Opioid Use Disorder</td>
<td></td>
</tr>
</tbody>
</table>
A.2 Search Strategies

Search Strategies

Database: Ovid MEDLINE(R) and Epub Ahead of Print, In-Process, In-Data-Review & Other Non-Indexed Citations, Daily and Versions <1946 to July 19, 2023>

Search Strategy:

1 Artificial Intelligence/ (39114)
2 exp Machine Learning/ (58584)
3 supervised machine learning/ (1665)
4 random forest/ (257)
5 random forest.tw,kf. (19867)
6 artificial intelligence.tw,kf. (36926)
7 machine learning.tw,kf. (91231)
8 computer-assisted.tw,kf. (30335)
9 AI.tw,kf. (45787)
10 neural networks, computer/ (48054)
11 neural network*.tw,kf. (97499)
12 deep learning/ (16019)
13 deep learning.tw,kf. (49663)
14 Support Vector Machine/ (10048)
15 support vector machine.tw,kf. (20661)
16 naive bayes.tw,kf. (3144)
17 k-nearest neighbor.tw,kf. (3713)
18 "boosting and bagging".tw,kf. (42)
19 gradient boosting machine*.tw,kf. (987)
20 predict* algorithm*.tw,kf. (5072)
21 predict* model*.tw,kf. (68103)
22 forecasting model*.tw,kf. (1549)
23 Substance-Related Disorders/ (105700)
24 amphetamine-related disorders/ (3499)
25 cocaine-related disorders/ (8974)
26 Opioid-Related Disorders/ (21687)
27 marijuana abuse/ (7064)
28 heroin dependence/ (9323)
29 morphine dependence/ (3556)
30 opium dependence/ (77)
31 substance abuse, intravenous/ (16803)
32 substance abuse, oral/ (8)
33 ((cocaine or substance or opioid or opiate or heroin or morphine oramphetamine or marijuana or fentanyl) adj1 (disorder* or abuse or "use" or dependence or addiction)).tw,kf. (113652)
34 Addiction Medicine/ (235)
35 (addiction* adj (medicine or psychiatry)).tw,kf. (1145)
36 Drug Overdose/ (14422)
37 Opiate Overdose/ (714)
38 overdos*.tw,kf. (29671)
Database: Embase <1974 to 2023 July 18>
Search Strategy:

1 artificial intelligence/ (59454)
2 exp machine learning/ (402356)
3 supervised machine learning/ (4240)
4 random forest/ (23487)
5 random forest.tw,kf. (24471)
6 artificial intelligence.tw,kf. (43400)
7 machine learning.tw,kf. (106507)
8 computer-assisted.tw,kf. (38801)
9 Al.tw,kf. (61268)
10 artificial neural network/ (50503)
11 neural network*.tw,kf. (116141)
12 deep learning/ (41509)
13 deep learning.tw,kf. (56886)
14 support vector machine/ (38275)
15 support vector machine.tw,kf. (25093)
16 naive bayes.tw,kf. (3852)
17 k-nearest neighbour.tw,kf. (4557)
18 "boosting and bagging".tw,kf. (49)
19 gradient boosting machine*.tw,kf. (1173)
20 predict* algorithm*.tw,kf. (7309)
21 predict* model*.tw,kf. (90467)
22 forecasting model*.tw,kf. (1703)
23 drug dependence/ (68489)
24 amphetamine dependence/ (756)
25 cocaine dependence/ (14232)
26 opiate addiction/ (28869)
27 cannabis addiction/ (11794)
28 heroin dependence/ (10006)
29 morphine addiction/ (3302)
30 opiate addiction/ (28869)
31 ((cocaine or substance or opioid or opiate or heroin or morphine or amphetamine or marijuana or fentanyl) adj1 (disorder* or abuse or "use" or dependence or addiction)).tw,kf. (153457)
32 Addiction Medicine/ (722)
33 (addiction* adj (medicine or psychiatry)).tw,kf. (2630)
34 Drug Overdose/ (35318)
35 Opiate Overdose/ (413)
36 overdos*.tw,kf. (42994)
37 or/1-22 (644481)
38 or/23-36 (272109)
39 37 and 38 (3093)
<table>
<thead>
<tr>
<th></th>
<th>Search Term</th>
<th>Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>(MH &quot;Artificial Intelligence+&quot;)</td>
<td>29,729</td>
</tr>
<tr>
<td>S2</td>
<td>(MH &quot;Machine Learning&quot;) OR (MH &quot;Deep Learning&quot;) OR (MH &quot;Support Vector Machine&quot;)</td>
<td>6,435</td>
</tr>
<tr>
<td>S3</td>
<td>(MH &quot;Random Forest&quot;)</td>
<td>474</td>
</tr>
<tr>
<td>S4</td>
<td>(MH &quot;Neural Networks (Computer)&quot;)</td>
<td>3,848</td>
</tr>
<tr>
<td>S5</td>
<td>TI &quot;artificial intelligence&quot; OR AB &quot;artificial intelligence&quot;</td>
<td>8,243</td>
</tr>
<tr>
<td>S6</td>
<td>TI &quot;machine learning&quot; OR AB &quot;machine learning&quot;</td>
<td>12,837</td>
</tr>
<tr>
<td>S7</td>
<td>TI &quot;random forest&quot; OR AB &quot;random forest&quot;</td>
<td>2,626</td>
</tr>
<tr>
<td>S8</td>
<td>TI &quot;computer-assisted&quot; OR AB &quot;computer-assisted&quot;</td>
<td>6,528</td>
</tr>
<tr>
<td>S9</td>
<td>TI &quot;AI&quot; OR AB &quot;AI&quot;</td>
<td>9,782</td>
</tr>
<tr>
<td>S10</td>
<td>TI &quot;neural network*&quot; OR AB &quot;neural network*&quot;</td>
<td>7,350</td>
</tr>
<tr>
<td>S11</td>
<td>TI &quot;deep learning&quot; OR AB &quot;deep learning&quot;</td>
<td>5,339</td>
</tr>
<tr>
<td>S12</td>
<td>TI &quot;support vector machine*&quot; OR AB &quot;support vector machine*&quot;</td>
<td>2,853</td>
</tr>
<tr>
<td>S13</td>
<td>TI &quot;naive bayes&quot; OR AB &quot;naive bayes&quot;</td>
<td>218</td>
</tr>
<tr>
<td>S14</td>
<td>TI &quot;k-nearest neighbour&quot; OR AB &quot;k-nearest neighbour&quot;</td>
<td>58</td>
</tr>
<tr>
<td>S15</td>
<td>TI (&quot;boosting and bagging&quot;) OR AB (&quot;boosting and bagging&quot;)</td>
<td>4</td>
</tr>
<tr>
<td>S16</td>
<td>TI &quot;gradient boosting machine*&quot; OR AB &quot;gradient boosting machine*&quot;</td>
<td>142</td>
</tr>
<tr>
<td>S17</td>
<td>TI &quot;predict* algorithm*&quot; OR AB &quot;predict* algorithm*&quot;</td>
<td>785</td>
</tr>
<tr>
<td>S18</td>
<td>TI &quot;predict* model*&quot; OR AB &quot;predict* model*&quot;</td>
<td>15,055</td>
</tr>
<tr>
<td>S19</td>
<td>TI &quot;forecasting model*&quot; OR AB &quot;forecasting model*&quot;</td>
<td>180</td>
</tr>
<tr>
<td>S20</td>
<td>S1ORS2ORS3ORS4 ORS5ORS6ORS7OR S8ORS9ORS10OR S11ORS12ORS13OR S14ORS15ORS16OR S17 OR S18 OR S19</td>
<td>73,636</td>
</tr>
<tr>
<td>S21</td>
<td>(MH &quot;Substance Dependence&quot;)</td>
<td>10,699</td>
</tr>
<tr>
<td>S22</td>
<td>TI (((cocaine or substance or opioid or opiate or heroin or morphine or amphetamine or marijuana or fentanyl) N1 (disorder* or abuse or &quot;use&quot; or dependence or addiction)) OR AB (((cocaine or substance or opioid or opiate or heroin or morphine or amphetamine or marijuana or fentanyl) N1 (disorder* or abuse or &quot;use&quot; or dependence or addiction)))</td>
<td>65,762</td>
</tr>
</tbody>
</table>
or heroin or morphine or amphetamine or marijuana or fentanyl) N1 (disorder* or abuse or "use" or dependence or addiction)) )

S23: TI ( (addiction* N1 (medicine or psychiatry or psychology)) ) AND AB ( (addiction* N1 (medicine or psychiatry or psychology)) )

115

S24: S21 OR S22 OR S23

73,615

S25: S20 AND S24

699

Database – APA PsycINFO
Interface - EBSCOhost Research Databases
Search Screen - Advanced Search
Total records = 1, 130


2,001

S2: (TI "Artificial Intelligence" OR AB "Artificial Intelligence" OR KW "Artificial Intelligence"

8,080

S3: TI "Machine Learning" OR AB "Machine Learning" OR KW "Machine Learning"

12,140

S4: TI "Neural Network*" OR AB "Neural Network*" OR KW "Neural Network*"

23,067

S5: TI "computer-assisted" OR AB "computer-assisted" OR KW "computer-assisted"

7,912

S6: TI “AI” OR AB “AI” OR KW “AI”

6,890

S7: TI deep learning in education OR AB “deep learning” OR KW “deep learning”

3,401

S8: TI “support vector machine” OR AB “support vector machine” OR KW “support vector machine”

2,581

S9: TI “naive bayes” OR AB “naive bayes” OR KW “naive bayes”

237

S10: TI “k-nearest neighbo?r” OR AB “k-nearest neighbo?r” OR KW “k-nearest neighbo?r”

71

S11: TI ( "boosting and bagging" ) OR AB ( "boosting and bagging" ) OR KW ( "boosting and bagging" )

10

S12: TI “gradient boosting machine*” OR AB “gradient boosting machine*” OR KW “gradient boosting machine*”

43

S13: TI “predict* algorithm*” OR AB “predict* algorithm*”

333

S14: TI “predict* model*” OR AB “predict* model*” OR KW “predict* model*”

7,415

S15: TI “forecasting model*” OR AB “forecasting model*” OR KW “forecasting model*”

242

S16: S1ORS2ORS3ORS4 ORS5ORS6ORS7OR S8ORS9ORS10OR S11ORS12ORS13OR S14 OR S15

61,458
| S17 | (((DE "Drug Addiction" OR DE "Drug Withdrawal" OR DE "Addiction Medicine" OR DE "Addiction Psychiatry" OR DE "Drug Dependency" OR DE "Drug Overdoses" OR DE "Intravenous Drug Usage" OR DE "Substance Use Disorder") OR (DE "Morphine Dependence")) OR (DE "Cannabis Use Disorder")) OR (DE "Substance Related and Addictive Disorders") OR (DE "Heroin Use Disorder") | 49,967 |
| S18 | TI ( ((cocaine or substance or opioid or opiate or heroin or morphine or amphetamine or marijuana or fentanyl) N1 (disorder* or abuse or "use" or dependence or addiction)). ) OR AB ( ((cocaine or substance or opioid or opiate or heroin or morphine or amphetamine or marijuana or fentanyl) N1 (disorder* or abuse or "use" or dependence or addiction)). ) OR KW ( ((cocaine or substance or opioid or opiate or heroin or morphine or amphetamine or marijuana or fentanyl) N1 (disorder* or abuse or "use" or dependence or addiction))) | 104,578 |
| S19 | TI ( (addiction* N1 (medicine or psychiatry)) ) OR AB ( (addiction* N1 (medicine or psychiatry)) ) OR KW ( (addiction* N1 (medicine or psychiatry))) | 958 |
| S20 | S17 OR S18 OR S19 | 130,991 |
| S21 | S16 AND S20 | 1,130 |

**Database: Web of Science Core Collection (University of British Columbia)**
Editions =Science Citation Index Expanded (SCI-EXPANDED)--1900-present; Social Sciences Citation Index; (SSCI)--1956-present; Arts & Humanities Citation Index (AHCI)--1975-present; Conference Proceedings Citation Index – Science (CPCI-S)--1990-present

**1,958 results from Web of Science Core Collection run July 19, 2023:**

1. TS= ("random forest" OR "artificial intelligence" OR "Machine Learning" OR "computer-assisted" OR "AI" OR "neural network*" OR "deep learning" OR "support vector machine" OR "naive bayes" OR "k-nearest neighbor" OR "boosting and bagging" OR "gradient boosting machine" OR "predict* algorithm*" OR "predict* model*" OR "forecasting model*")

1, 325, 616

2. TS=((cocaine or substance or opioid or opiate or heroin or morphine or amphetamine or marijuana or fentanyl) NEAR/1 (disorder* or abuse or "use" or dependence or addiction))

164, 543

3. TS= (addiction* NEAR/1 (medicine or psychiatry))

1,300

4. #2 or #3

165, 157

5. #4 AND #5
1,958
Appendix B  Chapter 3

B.1  Search Strategies

Database: Ovid MEDLINE(R) and Epub Ahead of Print, In-Process, In-Data-Review & Other Non-Indexed Citations, Daily and Versions <1946 to May 25, 2023>

Search Strategy:
1 Decision Support Systems, Clinical/ (9421)
2 decision support techniques/ (22449)
3 (clinic* decision adj2 (technique* or tool* or aid* or model* or system* or program* or software or rule* or interface or alert*)).tw,kf. (7998)
4 (computer* decision adj2 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)).tw,kf. (690)
5 (CDS adj1 (technique* or tool* or aid* or model* or system* or program* or software* or interface or alert*)).tw,kf. (934)
6 (decision adj2 tool*).tw,kf. (7898)
7 CDSS.tw,kf. (2460)
8 non-knowledge based.tw,kf. (9)
9 random forest/ (198)
10 random forest.tw,kf. (19026)
11 Decision Making, Computer-Assisted/ (2876)
12 Drug Therapy, Computer-Assisted/ (1691)
13 Clinical Decision Rules/ (911)
14 computer-assisted.tw,kf. (30144)
15 exp Substance-Related Disorders/ (309609)
16 Opioid-Related Disorders/ (21457)
17 ((cocaine or nicotine or alcohol or tobacco or substance or opioid or heroin or morphine or amphetamine or marijuana or drug*) adj1 (disorder* or abuse or "use" or dependence or addiction)).tw,kf. (256352)
18 exp Suicide/ (74249)
19 Schizophrenia/ (110771)
20 Anxiety Disorders/ (41250)
21 Attention Deficit Disorder with Hyperactivity/ (34504)
22 bipolar disorder/ (44996)
23 exp depressive disorder/ (121455)
24 mood disorders/ (15871)
25 stress disorders, traumatic/ or stress disorders, post-traumatic/ or stress disorders, traumatic, acute/ (41890)
Database: Embase <1974 to 2023 May 25>

Search Strategy:
1 clinical decision support system/ (5757)
2 decision support system/ (27473)
3 (clinic* decision adj2 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)).tw,kf. (11491)
4 (computer* decision adj2 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)).tw,kf. (910)
5 (CDS adj1 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)).tw,kf. (1232)
6 (decision adj2 tool*).tw,kf. (11555)
7 CDSS.tw,kf. (3141)
8 non-knowledge based.tw,kf. (13)
9 random forest.tw,kf. (24387)
10 random forest/ (23655)
11 computer assisted drug therapy/ (937)
12 clinical decision rule/ (698)
13 computer-assisted.tw,kf. (38895)
14 exp drug dependence/ (274587)
15 opiate addiction/ or narcotic dependence/ (30489)
16 ((cocaine or nicotine or alcohol or tobacco or substance or opioid or heroin or morphine or amphetamine or marijuana or drug) adj2 (disorder* or abuse or "use" or addiction or dependence)).tw,kf. (392999)
17 suicide/ or suicidal behavior/ (79821)
18 psychosis/ (105202)
19 schizophrenia/ or schizophrenia spectrum disorder/ (195193)
20 anxiety disorder/ (94251)
21 attention deficit hyperactivity disorder/ (8239)
bipolar disorder/ (65865)
depression/ (474254)
mood disorder/ (52584)
posttraumatic stress disorder/ (79252)
personality disorder/ (31476)
(suicide or schizo* or anxiety or Attention Deficit or post-traumatic or ADHD or bipolar or depress* or mood disorder* or stress disorder* or Personality Disorder* or psychosis).tw,kf. (1387005)
or/1-13 (116473)
or/14-27 (2034057)
28 and 29 (6740)
limit 30 to english language (6536)

Database: Web of Science Core Collection (University of British Columbia)
Editions = A&HCI, ESCI, CPCI-SSH, CPCI-S, SCI-EXPANDED, SSCI

<table>
<thead>
<tr>
<th>Line</th>
<th>Search Query</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TS=(&quot;clinic* decision&quot; NEAR/2 (technique* or tool* or aid* or model* or system* or program* or software or rule* or interface or alert*))</td>
<td>10023</td>
</tr>
<tr>
<td>2</td>
<td>TS=&quot;computer decision&quot; NEAR/2 (technique* or tool* or aid* or model* or system* or program* or software or rule* or interface or alert*))</td>
<td>180</td>
</tr>
<tr>
<td>3</td>
<td>TS=&quot;CDSS&quot; NEAR/2 (technique* or tool* or aid* or model* or system* or program*)</td>
<td>1244</td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td>Term Count</td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------------------------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>4</td>
<td>TS=(decision NEAR/2 tool*)</td>
<td>27230</td>
</tr>
<tr>
<td>5</td>
<td>TS=“non-knowledge based”</td>
<td>68</td>
</tr>
<tr>
<td>6</td>
<td>TS=“random forest”</td>
<td>72063</td>
</tr>
<tr>
<td>7</td>
<td>TS=“computer-assisted”</td>
<td>46024</td>
</tr>
<tr>
<td>8</td>
<td>TS=“computer assisted”</td>
<td>66382</td>
</tr>
<tr>
<td>9</td>
<td>TS=((cocaine or nicotine or alcohol or tobacco or substance or opioid or heroin or morphine or heroin or morphine or amphetamine or marijuana or drug*) NEAR/1 (disorder* or abuse or &quot;use&quot; or dependence or addiction))</td>
<td>344856</td>
</tr>
<tr>
<td>10</td>
<td>TS=(suicide or schizo* or anxiety or post-traumatic or Attention Deficit or ADHD or bipolar or depress* or mood disorder* or stress disorder* or “Personality Disorder*” or psychosis)</td>
<td>1606778</td>
</tr>
<tr>
<td>11</td>
<td>#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8</td>
<td>172895</td>
</tr>
<tr>
<td>12</td>
<td>#9 OR #10</td>
<td>186937</td>
</tr>
<tr>
<td></td>
<td>Query</td>
<td>Results</td>
</tr>
<tr>
<td>---</td>
<td>----------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>S1</td>
<td>DE &quot;Decision Support Systems&quot;</td>
<td>3,725</td>
</tr>
<tr>
<td>S2</td>
<td>TI ( (clinic* decision support N2 (technique* or tool* or aid* or model* or system* or program* or software or rule* or interface or alert*)) ) OR AB ( (clinic* decision support N2 (technique* or tool* or aid* or model* or system* or program* or software or rule* or interface or alert*)) )</td>
<td>411</td>
</tr>
<tr>
<td>S3</td>
<td>( (computer* decision support N2 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)) ) OR ( (computer* decision support N2 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)) )</td>
<td>169</td>
</tr>
<tr>
<td>S4</td>
<td>TI ( ((CDS N1 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)) ) OR AB ( (CDS N1 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)) )</td>
<td>98</td>
</tr>
<tr>
<td>S5</td>
<td>TI (decision N2 tool*) OR AB (decision N2 tool*)</td>
<td>1,762</td>
</tr>
<tr>
<td></td>
<td>Term</td>
<td>Count</td>
</tr>
<tr>
<td>---</td>
<td>----------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>S6</td>
<td>TI CDSS OR AB CDSS</td>
<td>369</td>
</tr>
<tr>
<td>S7</td>
<td>TI non-knowledge based OR AB non-knowledge based</td>
<td>1</td>
</tr>
<tr>
<td>S8</td>
<td>TI random forest OR AB random forest</td>
<td>1,482</td>
</tr>
<tr>
<td>S9</td>
<td>(DE &quot;Computer Assisted Therapy&quot;)</td>
<td>1,221</td>
</tr>
<tr>
<td>S10</td>
<td>TI computer-assisted OR AB computer-assisted</td>
<td>6,681</td>
</tr>
<tr>
<td>S11</td>
<td>DE &quot;Substance Use Disorder&quot; OR DE &quot;Drug Addiction&quot; OR DE &quot;Substance Related and Addictive Disorders&quot;</td>
<td>22,778</td>
</tr>
<tr>
<td>S12</td>
<td>TI ( (cocaine or nicotine or alcohol or tobacco or substance or opioid or heroin or morphine or amphetamine or marijuana or substance or drug*) N1 (disorder* or abuse or &quot;use&quot; or dependence or addiction) ) OR AB ( (cocaine or nicotine or alcohol or tobacco or substance or opioid or heroin or morphine or amphetamine or marijuana or substance or drug*) N1 (disorder* or abuse or &quot;use&quot; or dependence or addiction) )</td>
<td>190,032</td>
</tr>
<tr>
<td>S13</td>
<td>DE &quot;Suicide&quot; OR DE &quot;Suicidal Behavior&quot; OR DE &quot;Suicidal Ideation&quot;</td>
<td>44,542</td>
</tr>
<tr>
<td>S14</td>
<td>DE &quot;Schizophrenia&quot;</td>
<td>105,773</td>
</tr>
<tr>
<td>S15</td>
<td>DE &quot;Anxiety&quot; OR DE &quot;Anxiety Disorders&quot;</td>
<td>122,282</td>
</tr>
<tr>
<td>S16</td>
<td>DE &quot;Attention Deficit Disorder with Hyperactivity&quot; OR DE &quot;Attention Deficit Disorder&quot;</td>
<td>35,582</td>
</tr>
<tr>
<td>S17</td>
<td>DE &quot;Bipolar Disorder&quot;</td>
<td>38,730</td>
</tr>
<tr>
<td>S18</td>
<td>DE &quot;Major Depression&quot;</td>
<td>148,620</td>
</tr>
<tr>
<td>S19</td>
<td>DE &quot;Affective Disorders&quot;</td>
<td>15,868</td>
</tr>
<tr>
<td>S20</td>
<td>DE &quot;Posttraumatic Stress&quot; OR DE &quot;Posttraumatic Stress Disorder&quot; OR DE &quot;Stress and Trauma Related Disorders&quot;</td>
<td>40,797</td>
</tr>
<tr>
<td>S21</td>
<td>DE &quot;Personality Disorders&quot;</td>
<td>21,605</td>
</tr>
<tr>
<td>S22</td>
<td>TI ( (suicide or schizo* or anxiety or posttraumatic or post-traumatic or Attention Deficit or...</td>
<td>753,643</td>
</tr>
</tbody>
</table>
ADHD or bipolar or depress* or mood disorder* or stress disorder* or Personality Disorder* or psychosis) OR AB ( (suicide or schizo* or anxiety or posttraumatic or post-traumatic or Attention Deficit or ADHD or bipolar or depress* or mood disorder* or stress disorder* or Personality Disorder* or psychosis) )

| S23  | S1ORS2ORS3ORS4 ORS5ORS6ORS7OR S8ORS9ORS10 | 14,864  |
| S24  | S11ORS12ORS13OR S14ORS15ORS16OR S17ORS18ORS19OR S20 OR S21 OR S22 | 939,438 |
| S25  | S23 AND S24 | 2,535  |
| S26  | Narrow by Language: - english | 2,431  |

**Database** – CINAHL

**Interface** - EBSCOhost Research Databases

**Search Screen** - Advanced Search

Total records = 2144

<table>
<thead>
<tr>
<th>#</th>
<th>Query</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>(MH &quot;Decision Support Systems, Clinical&quot;)</td>
<td>6,445</td>
</tr>
<tr>
<td>S2</td>
<td>(MH &quot;Decision Support Techniques&quot;)</td>
<td>7,707</td>
</tr>
<tr>
<td>S3</td>
<td>TI ( (clinic* decision support N2 (technique* or tool* or aid* or model* or system* or program* or software or rule* or interface or alert*)) ) OR AB ( (clinic* decision support N2 (technique* or tool* or aid* or model* or system* or program* or software or rule* or interface or alert*)) )</td>
<td>2,747</td>
</tr>
<tr>
<td>S4</td>
<td>TI ( (computer* decision support N2 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)) ) OR ( (computer* decision support N2 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)) ) OR AB ( (computer* decision support N2 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)) ) ) OR ( (computer* decision support N2 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)) ) ) OR AB ( (computer* decision support N2 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)) ) ) OR ( (computer* decision support N2 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)) )</td>
<td>709</td>
</tr>
<tr>
<td>S5</td>
<td>TI ( (CDS N1 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)) ) OR AB ( (CDS N1 (technique* or tool* or aid* or model* or system* or program* or software* or rule* or interface or alert*)) )</td>
<td>386</td>
</tr>
<tr>
<td>S6</td>
<td>TI (decision N2 tool*) OR AB (decision N2 tool*)</td>
<td>3,789</td>
</tr>
<tr>
<td>S7</td>
<td>TI CDSS OR AB CDSS</td>
<td>701</td>
</tr>
<tr>
<td>S8</td>
<td>TI non-knowledge based OR AB non-knowledge based</td>
<td>2</td>
</tr>
<tr>
<td>S9</td>
<td>(MH &quot;Random Forest&quot;)</td>
<td>438</td>
</tr>
<tr>
<td>S10</td>
<td>TI random forest OR AB random forest</td>
<td>3,086</td>
</tr>
<tr>
<td>S11</td>
<td>(MH &quot;Decision Making, Computer Assisted&quot;)</td>
<td>1,406</td>
</tr>
<tr>
<td>S12</td>
<td>(MH &quot;Drug Therapy, Computer Assisted&quot;)</td>
<td>490</td>
</tr>
<tr>
<td>S13</td>
<td>TI computer-assisted OR AB computer-assisted</td>
<td>6,536</td>
</tr>
<tr>
<td>S14</td>
<td>(MH &quot;Substance Use Disorders+&quot;)</td>
<td>184,527</td>
</tr>
<tr>
<td>S15</td>
<td>TI ( (cocaine or nicotine or alcohol or tobacco or substance)</td>
<td>131,342</td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td>Count</td>
</tr>
<tr>
<td>---</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>S16</td>
<td>(MH &quot;Suicide+&quot;) OR (MH &quot;Suicidal Ideation&quot;)</td>
<td>37,424</td>
</tr>
<tr>
<td>S17</td>
<td>(MH &quot;Schizophrenia+&quot;)</td>
<td>29,104</td>
</tr>
<tr>
<td>S18</td>
<td>(MH &quot;Anxiety+&quot;) OR (MH &quot;Anxiety Disorders+&quot;)</td>
<td>105,047</td>
</tr>
<tr>
<td>S19</td>
<td>(MH &quot;Attention Deficit Hyperactivity Disorder&quot;)</td>
<td>18,876</td>
</tr>
<tr>
<td>S20</td>
<td>(MH &quot;Bipolar Disorder+&quot;)</td>
<td>13,418</td>
</tr>
<tr>
<td>S21</td>
<td>(MH &quot;Depression&quot;)</td>
<td>128,089</td>
</tr>
<tr>
<td>S22</td>
<td>(MH &quot;Affective Disorders&quot;)</td>
<td>8,722</td>
</tr>
<tr>
<td>S23</td>
<td>(MH &quot;Stress Disorders, Post-Traumatic&quot;)</td>
<td>26,757</td>
</tr>
<tr>
<td>S24</td>
<td>(MH &quot;Personality Disorders&quot;)</td>
<td>5,097</td>
</tr>
<tr>
<td>S25</td>
<td>TI ( suicide or schizo* or anxiety or posttraumatic or posttraumatic or Attention Deficit or ADHD or bipolar or depress* or mood disorder* or stress disorder* or Personality Disorder* or psychosis ) OR AB ( suicide or schizo* or anxiety or posttraumatic or posttraumatic or Attention Deficit or ADHD or bipolar or depress* or mood disorder* or stress disorder* or Personality Disorder* or psychosis )</td>
<td>349,115</td>
</tr>
<tr>
<td>S26</td>
<td>S1 OR S2 OR S3 OR S4 OR S5 ORS6 OR S7 S8 OR S9 OR S10 OR S11 OR S12 OR S13</td>
<td>631,882</td>
</tr>
<tr>
<td>S27</td>
<td>S14 OR S15 OR S16 S17 OR S18 OR S19 S20 OR S21 OR S22 S23 OR S24 OR S25</td>
<td>631,882</td>
</tr>
<tr>
<td>S28</td>
<td>S26 AND S27</td>
<td>2,150</td>
</tr>
<tr>
<td>-----</td>
<td>-------------</td>
<td>-------</td>
</tr>
<tr>
<td>S29</td>
<td>Narrow by Language: - english</td>
<td>2,144</td>
</tr>
</tbody>
</table>
Appendix C  Chapter 4 Appendix

C.1  British Columbia Provincial Overdose Cohort Descriptions for Data Sources

BC Coroner's Service (BCCS): BCCS conducts investigations into all unexpected and undetermined deaths from illicit drug overdoses in BC. The data provided by BCCS includes details such as the date and location of death or injury, demographic information about the deceased, the setting of the death, the individual's residence, substance use and method of intake, along with post-mortem toxicological findings.

BC Emergency Health Services (BCEHS): BCEHS data comprises details regarding the timing and place of overdose incidents, demographic profiles of the patients, and information gathered from the emergency response, including the paramedics' assessments, interventions, and patient transport specifics, like the administration of naloxone. Overdose events are identified using impression codes from the BCEHS Patient Care Report (PCR) data, which include various code combinations employed by first responders. Initially sourced from paper-based reports by paramedics, this data is now largely entered into the electronic SIREN system since April 2019, with new codes for overdose incidents. The 2017 Cohort encompasses data from both the PCIS and SIREN systems.

Enhanced Emergency Department (EED) Records: These records entail manual reporting of opioid- and drug-related overdoses from emergency departments across three BC Health Authorities: the Interior Health Authority (IHA), Vancouver Island Health Authority (VIHA), and Northern Health Authority (NHA).

Drug and Poison Information Centre (DPIC): DPIC data logs inquiries related to poisoning, encompassing both public and medical consultations. It captures the time of the call,
demographics of the patient involved, originating location of the call, implicated substances, method of substance use, exhibited symptoms, and the eventual outcome of the case.

National Ambulatory Care Reporting System (NACRS): Managed by the Canadian Institute for Health Information (CIHI), NACRS is a comprehensive database that records patient visits to ambulatory care services, both hospital and community-based. The data, which is collected in real-time during service provision, includes demographic, clinical, administrative, financial, and specific service details. It also captures information about patient discharges, deaths, and transfers within a fiscal year. As of 2016, NACRS covered 67% of all ED visits in BC, with near-total inclusion of visits from urban EDs, particularly within the Vancouver Coastal Health Authority and Fraser Health Authority.

Medical Services Plan (MSP): The MSP database houses records of all fee-for-service medical consultations billed to the provincial health insurance program, excluding those under alternative payment schemes. It categorizes providers into physicians and allied health professionals like physiotherapists. Each visit record details the patient's demographic data, date of service, location, provider category, and the nature of the visit, encoded using the ICD-9 classification system.

Discharge Abstract Database (DAD): DAD encompasses detailed records of patient exits, whether through discharge, transfer, or death, from acute care hospitals across BC. The data includes demographic specifics of the patient, duration and setting of the hospital stay, attending healthcare provider categories, along with diagnoses and procedures performed, coded via the ICD-10 system.
PharmaNet (PNET): PharmaNet logs every prescription dispensed in outpatient settings throughout BC, capturing patient demographic information, the dispensing location, prescriber type, and detailed medication data including the drug name, formulation, and prescribed dosage.

Chronic Disease Registry (CDR): The BC Chronic Disease Registries (CDR), established in the 1990s and updated annually, serve as a comprehensive source for measuring the burden of chronic diseases across the province. They track the incidence and prevalence of conditions over time and scrutinize disparities in disease burden by sex, age, and region. Plans are underway to extend this analysis to include indigenous identity and socioeconomic status, enhancing the understanding of health inequalities.

The CDR is a suite of 25 registries curated by the Office of the Provincial Health Officer (OPHO), with new registries added periodically. These registries are informed by data from various administrative sources, including the Medical Services Plan, Discharge Abstract Databases, PharmaNet, and Client Roster.

The data harvested from the CDR enables researchers to estimate both the incidence and the lifetime prevalence of 25 chronic conditions in BC. It also allows for the calculation of active healthcare contact prevalence for 11 relapsing-remitting diseases. These include chronic respiratory diseases like asthma and COPD; cardiovascular diseases such as acute myocardial infarction, heart failure, and various forms of stroke; neurological disorders including Alzheimer’s, epilepsy, and Parkinson’s disease; mental and substance use disorders like depression, anxiety, and schizophrenia; musculoskeletal disorders such as gout and arthritis; as well as diabetes and kidney diseases.
### CDR Case Definitions

Table 24 CDR Case Definitions (338)

<table>
<thead>
<tr>
<th>Group</th>
<th>Disease</th>
<th>BCCDR Algorithm</th>
<th>Validation Study Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circulatory Diseases</td>
<td>Ischemic heart disease (20+)</td>
<td>(2+ P for Angina +1Rx in 1 year) or (1+ Angina SP + 1Rx)</td>
<td>(339,340)</td>
</tr>
<tr>
<td></td>
<td>Acute myocardial infarction (20+) #</td>
<td>1+ H ever</td>
<td>(341)</td>
</tr>
<tr>
<td></td>
<td>Heart failure (1+)</td>
<td>1+ H or 2+ P within 1 year</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>High blood pressure (hypertension, 20+)</td>
<td>1+ H or 2+ P within 2 years</td>
<td>(342)</td>
</tr>
<tr>
<td></td>
<td>Hospitalized stroke (20+)</td>
<td>1+ H ever</td>
<td>(343–345)</td>
</tr>
<tr>
<td></td>
<td>Hospitalized haemorrhagic stroke (20+)</td>
<td>1+ H ever</td>
<td>(346)</td>
</tr>
<tr>
<td></td>
<td>Hospitalized ischemic stroke (20+)</td>
<td>1+ H ever</td>
<td>(346)</td>
</tr>
<tr>
<td></td>
<td>Hospitalized transient ischemic attack (TIA, 20+)</td>
<td>1+ H ever</td>
<td>(346)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Diabetes mellitus (1+)</td>
<td>1+H ever or 2+P in 1 year or</td>
<td>(347)</td>
</tr>
<tr>
<td>Inflammatory</td>
<td>Rheumatoid arthritis (1+)</td>
<td>2+ P (61-720 days apart)</td>
<td>(348–350)</td>
</tr>
<tr>
<td></td>
<td>Juvenile idiopathic arthritis (0-15)</td>
<td>1+H or 2+ P (8+ weeks apart) in 2 years</td>
<td>(351)</td>
</tr>
<tr>
<td>Kidney Diseases</td>
<td>Chronic kidney disease (1+)</td>
<td>1+ H or 2+ P in 1 year</td>
<td>(352)</td>
</tr>
<tr>
<td>Category</td>
<td>Condition</td>
<td>Criteria</td>
<td>Code(s)</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Mental Health</td>
<td>Mood and Anxiety disorders (1+)</td>
<td>1 H or 2P within 1 year</td>
<td>(353)</td>
</tr>
<tr>
<td></td>
<td>Depressive disorders</td>
<td>1H or 2P in 1Y</td>
<td>(354,355)</td>
</tr>
<tr>
<td></td>
<td>Schizophrenia &amp; delusional disorders (10+)</td>
<td>1H or 2P (30+ days apart) within 2 years</td>
<td>(356,357)</td>
</tr>
<tr>
<td>Musculoskeletal Disorders</td>
<td>Gout (20+)</td>
<td>1+ H or 2+ P (1+ day apart) within 5 years</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Osteoarthritis (1+)</td>
<td>1H or 2P in 1 year</td>
<td>(350,358)</td>
</tr>
<tr>
<td></td>
<td>Osteoporosis (50+)</td>
<td>1+ H or 2+ P or 2+ Rx in 1 year</td>
<td>(359,360)</td>
</tr>
<tr>
<td>Neurological Disorders</td>
<td>Alzheimer's and other dementias (40+)</td>
<td>1+H or 3+ P (30+ days apart) or 1+ Rx</td>
<td>(361,362)</td>
</tr>
<tr>
<td></td>
<td>Epilepsy (1+)</td>
<td>Aged 1-19 years: 3+ P (30+ days apart) within 2</td>
<td>(363)</td>
</tr>
<tr>
<td></td>
<td>Multiple sclerosis (20+)</td>
<td>1+ H or 5 + P within 2 years</td>
<td>(349,364)</td>
</tr>
<tr>
<td></td>
<td>Parkinson’s disease (40+)</td>
<td>2+ P (30+ days apart) within 1 year</td>
<td>(365)</td>
</tr>
<tr>
<td>Respiratory Diseases</td>
<td>Asthma (1+)</td>
<td>(1+ H or 2+ P within 1 year) or (1+ P and 2+ Rx within 1 year)</td>
<td>(366–368)</td>
</tr>
<tr>
<td></td>
<td>Chronic obstructive pulmonary disease (35+)</td>
<td>1+ H or 2+ P within 1 year</td>
<td>(369)</td>
</tr>
<tr>
<td>Substance Misuse</td>
<td>Substance Misuse Disorders (1+)</td>
<td>1+ H or 2+ P in 1 year</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Appendix D  Computational Tools and Code

D.1  Packages utilized on RStudio

Attached Packages:

1. arsenal, 3.6.2
2. MatchIt, 4.55
3. knn, 1.3.1
4. workflows, 0.2.1
5. tune, 0.1.1
6. rsample, 0.0.8
7. recipes, 0.1.14
8. parsnip, 0.1.4
9. modeldata, 0.1.0
10. infer, 0.5.3
11. dials, 0.0.9
12. scales, 1.2.1
13. broom, 1.0.4
14. tidymodels, 0.1.1
15. pROC, 1.18.2
16. here, 1.0.1
17. data.table, 1.14.2
18. mlbench, 2.1-3.1
19. randomForest, 4.6-14
20. SHAPforxgboost, 0.1.3
21. ROCR, 1.0-11
22. rnn, 1.9.0
23. imbalance, 1.0.2.1
24. ROSE, 0.0-4
25. Boruta, 8.0.0
26. yardstick, 0.0.7
27. naniar, 0.6.0
28. vctrs, 0.6.3
29. mice, 3.15.0
30. xgboost, 1.2.0.1
31. caret, 6.0-94
32. lattice, 0.20-41
33. gapminder, 0.3.0
34. lubridate, 1.9.2
35. forcats, 1.0.0
36. stringr, 1.5.0
37. dplyr, 1.1.2
38. purrr, 1.0.1
39. readr, 2.1.4
40. tidyr, 1.3.0
41. tibble, 3.2.1
42. ggplot2, 3.4.2
43. tidyverse, 2.0.0
44. rlang, 1.1.1
45. tensorflow, 2.2.0
46. keras, 2.3.0.0
47. stringi, 1.7.6
48. foreach, 1.5.2
49. checkmate, 2.0.0
50. lhs, 1.1.1
51. lava, 1.6.8.1
52. pkgconfig, 2.0.3
53. e1071, 1.7-8
54. pROC, 1.18.0
55. survival, 3.3-5
56. abind, 1.4-5
57. tseries, 0.10-50
58. timeDate, 3043.102
59. ModelMetrics, 1.2.2.2
60. recipes, 0.1.16
61. dplyr, 1.0.7
62. tidyr, 1.1.4
63. tibble, 3.1.6
64. ggplot2, 3.3.5
65. keras, 2.7.0
66. tensorflow, 2.7.0
67. coorplot, 0.84

Packages loaded via a namespace (and not attached):

1. colorspace, 2.0-3
2. ggsignif, 0.6.0
3. class, 7.3-17
4. rprojroot, 2.0.3
5. visdat, 0.5.3
6. base64enc, 0.1-3
7. rstudioapi, 0.14
8. listenv, 0.8.0
9. furrr, 0.2.1
10. ggpubr, 0.6.0
11. prodlim, 2019.11.13
12. farver, 2.1.0
13. fansi, 1.0.3
14. codetools, 0.2-18
15. splines, 4.0.5
16. knitr

D.2 Code for forest plots

All coding files were extracted via R Markdown (370). Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the Knit button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

Code for forest plots

```r
library(forestplot)

## Loading required package: grid
## Loading required package: checkmate
## Loading required package: abind

library(tidyverse)

## ── Attaching core tidyverse packages ────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.3 ✔ readr 2.1.4
## ✔ forcats   1.0.0 ✔ stringr 1.5.0
## ✔ ggplot2   3.4.1 ✔ tibble 3.2.1
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0
## ✔ purrr     1.0.2

## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag() masks stats::lag()
## ✔ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

# P <- AUCCIforestplot |> 
# ggplot(aes(y = reorder(Model_type, +AUC))) +
#  theme_classic() +
#  geom_point(aes(x= AUC), shape=15, size=3) +
#  geom_linerange(aes(xmin=lower, xmax=upper))+
#  labs(x="AUC scores", y="Model Type")

# p_left <-
# AUCCIforestplot %>%
#  ggplot(aes(y = reorder(Model_type, +AUC)))
# p_left
# #
# # p_left <-
```

258
```r
# p_left +
# geom_text(aes(x = 0, label = Model_type), hjust = 0, fontface = "bold")
#
# p_left <-
# p_left +
# geom_text(
#   aes(x = 1, label = AUCCI),
#   hjust = 0,
#   fontface = ifelse(AUCCIforestplot$AUCCI == "AUC (95% CI)", "bold", "plain")
# )
#
# p_left
#
# p_left <-
# p_left +
# theme_void() +
# coord_cartesian(xlim = c(0, 4))
#
# p_left

# P <- AUCCIFOD |>  
# ggplot(aes(y = reorder(Model_type, +AUC))) +
# theme_classic() +
# geom_point(aes(x = AUC), shape=15, size=3) +
# geom_linerange(aes(xmin=lower, xmax=upper)) +
# labs(x="AUC scores", y="Model Type")
#
# P

# p_left <-
# AUCCIFOD %>%
# ggplot(aes(y = reorder(Model_type, +AUC)))
# p_left
#
# p_left <-
# p_left +
# geom_text(aes(x = 0, label = Model_type), hjust = 0, fontface = "bold")
#
# p_left <-
# p_left +
# geom_text(
#   aes(x = 1, label = AUCCI),
#   hjust = 0,
#   fontface = ifelse(AUCCIFOD$AUCCI == "AUC (95% CI)", "bold", "plain")
# )
#
# p_left
#
# p_left <-
```
# p_left +
# theme_void() +
# coord_cartesian(xlim = c(0, 4))
#
# p_left
D.3 Data wrangling code

All coding files were extracted via R Markdown (370). Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see [http://rmarkdown.rstudio.com](http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

Total instances file

```
library(tidyverse)

## -- Attaching core tidyverse packages ---------------------------------- tidyverse 2.0.0 --
## v dplyr 1.1.2  v readr 2.1.4
## v forcats 1.0.0  v stringr 1.5.0
## v ggplot2 3.4.2  v tibble 3.2.1
## v lubridate 1.9.2  v tidyr 1.3.0
## v purrr 1.0.1
## -- Conflicts ---------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag() masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(dplyr)
library(gapminder)

## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)  
##   problems(dat)

# Pharmanetspss_total <- full_join(x = vw_od_case, y = Pharmanetspss, by = "moh_study_id", all.x = TRUE)
# MSP_clean_total <- full_join(x = vw_od_case, y = MSP_clean, by = "moh_study_id", all.x = TRUE)
# vw_dad_total <- full_join(x = vw_od_case, y = vw_dad, by = "moh_study_id", all.x = TRUE)

glimpse(Pharmanetspss)
## Rows: 57,902,314
## Columns: 7

glimpse(vw_dad)
## Rows: 1,663,940
## Columns: 141

glimpse(MSP_clean)
```

261
Icd Code 9 Extraction

```r
library(rlang)
library(tidyverse)

## -- Attaching core tidyverse packages ------------------------- tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts --------------------------------------------- tidyverse_conflicts() --
## x purrr::%@%()     masks rlang::%@%()
## x dplyr::filter()    masks stats::filter()
## x purrr::flatten()     masks rlang::flatten()
## x purrr::flatten_chr() masks rlang::flatten_chr()
## x purrr::flatten_dbl() masks rlang::flatten_dbl()
## x purrr::flatten_int() masks rlang::flatten_int()
## x purrr::flatten_lgl() masks rlang::flatten_lgl()
## x purrr::flatten_raw() masks rlang::flatten_raw()
## x purrr::invoke()    masks rlang::invoke()
## x dplyr::lag()       masks stats::lag()
## x purrr::splice()    masks rlang::splice()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(dplyr)
library(gapminder)
library(caret)

## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
## lift

library(xgboost)

##
## Attaching package: 'xgboost'
```
library(forcats)
library(mice)

library(vctrs)

MSP_clean <- read_csv("U:/Andy's Thesis/Raw_data_used/MSP_clean.csv", show_col_types = FALSE)
glimpse(MSP_clean)

ICD9_0 <- MSP_clean %>%
  mutate(across(starts_with("diag_cd "), ~case_when(
    (.x >= 2910) & (.x <= 2919) ~ 1,
    (.x >= 30300) & (.x <= 30303) ~ 1,
    (.x >= 3140) & (.x <= 3149) ~ 1,
    (.x >= 3140) & (.x <= 3149) ~ 1,
    .x == "291" ~ 1,
  
#testing

# TRUE ~ 0), .names = "derivde{.col}")))
#
# ICD9_1 <- (MSP_clean %>%
# mutate(across(starts_with("diag_cd_"), ~case_when(
#   (x >= 2910) & (x <= 2919) ~ 1,
#   (x >= 30300) & (x <= 30303) ~ 1,
#   (x >= 3140) & (x <= 3149) ~ 1,
#   (x >= 3140) & (x <= 3149) ~ 1,
#   x=="291"~ 1,
#   TRUE ~ 0), .names = "derivde2{.col}")))
#
# table(ICD9_0$derivdediag_cd_1)
# table(ICD9_1$derivde2diag_cd_1)
#
# Test1 <- ICD9_1 %>%
# select(moh_study_id, derivde2diag_cd_1, derivde2diag_cd_2, derivde2diag_cd_3)
#
# Test2 <- ICD9_0 %>%
# select(moh_study_id, derivdediag_cd_1, derivdediag_cd_2, derivdediag_cd_3)
#
# icd9_11 <- Test1 %>%
# group_by(moh_study_id) %>%
# summarize(derivde2diag_cd_1 = sum(derivde2diag_cd_1),
# derivde2diag_cd_2 = sum(derivde2diag_cd_2),
# derivde2diag_cd_3 = sum(derivde2diag_cd_3))
#
# icd9_12 < icd9_11 %>%
# mutate(Total = select(., moh_study_id) %>%
# rowSums(na.rm = TRUE))
#
# icd9_13 <- icd9_12 %>%
# mutate(earlyonsetdisorders = if_else(Total >= 1, 1, 0))
#
# icd9_14 <- icd9_13 %>%
# select(moh_study_id, earlyonsetdisorders)
#
# icd9_21 <- Test2 %>%
# group_by(moh_study_id) %>%
# summarize(derivdediag_cd_1 = sum(derivdediag_cd_1),
# derivdediag_cd_2 = sum(derivdediag_cd_2),
# derivdeddiag_cd_3 = sum(derivdeddiag_cd_3)

# icd9_22 <- icd9_21 %>%
#   mutate(Total = select(., moh_study_id) %>%
#         rowSums(na.rm = TRUE))

# icd9_23 <- icd9_22 %>%
#   mutate(earlyonsetdisorders = if_else(Total >= 1, 1, 0))

# icd9_24 <- icd9_23 %>%
#   select(moh_study_id, earlyonsetdisorders)

# table(icd9_14$earlyonsetdisorders)
# table(icd9_24$earlyonsetdisorders)

# Test3 <- merge(icd9_14, icd9_24, by = "moh_study_id", all.x = TRUE)

# Test4 <- Test3 %>% mutate(new = earlyonsetdisorders + earlyonsetdisorders)

# unique(MSP_clean$diag_cd_1)
# length(which(MSP_clean$diag_cd_1 == 291.0))
# length(which(MSP_clean$diag_cd_1 == 291))
# length(which(MSP_clean$diag_cd_1 == 2910))

###Template

# ICD9_0 <- (MSP_clean %>%
#   mutate(across(starts_with("diag_cd_"), ~case_when(.x == "E8500" ~ 1,
#           .x == "3040" ~ 1, .x == "3047" ~ 1,
#           .x == "3055" ~ 1, .x == "9650" ~ 1,
#           TRUE ~ 0), .names = "derivde{.col}")))

# ICD9_00 <- (MSP_clean %>%
#   mutate(across(starts_with("diag_cd_"), ~case_when((.x >= 301) & (.x <= 301.9) ~ 1,
#           TRUE ~ 0), .names = "derivde{.col}")))

###start

# ICD9_0 <- (MSP_clean %>%
#   mutate(across(starts_with("diag_cd_"), ~case_when(
#           .x == "1128" ~ 1,
#           .x == "11281" ~ 1,
#           .x == "421" ~ 1,
#           .x == "4210" ~ 1,
# .x== "4211" ~ 1,
# .x=="4219"~ 1,
# .x== "422" ~ 1,
# .x=="4220"~ 1,
# .x== "4229" ~ 1,
# .x=="42290"~ 1,
# .x== "42291" ~ 1,
# .x== "42292" ~ 1,
# .x=="42293" ~ 1,
# .x== "42299" ~ 1,
# .x== "424" ~ 1,
# .x=="4240"~ 1,
# .x== "4241" ~ 1,
# .x== "4242" ~ 1,
# .x=="4243" ~ 1,
# .x== "4249" ~ 1,
# .x== "42490" ~ 1,
# .x=="42491"~ 1,
# .x== "42499" ~ 1,
#
# TRUE ~ 0), .names = "derivde{.col}")

# ICD9_0 <- (MSP_clean %>%
# mutate(across(starts_with("diag_cd_"), ~case_when(
# .x=="038"~ 1,
# .x== "0380" ~ 1,
# .x=="0381"~ 1,
# .x=="03810"~ 1,
# .x== "03812" ~ 1,
# .x== "03819" ~ 1,
# .x== "03811" ~ 1,
# .x=="0382" ~ 1,
# .x== "0383" ~ 1,
# .x=="0384"~ 1,
# .x== "03840" ~ 1,
# .x== "03841" ~ 1,
# .x== "03842" ~ 1,
# .x== "03843" ~ 1,
# .x== "03844" ~ 1,
# .x== "03849"~ 1,
# .x== "0388" ~ 1,
# .x== "0389" ~ 1,
# .x=="4151"~ 1,
# .x== "41512" ~ 1,
# .x== "4229" ~ 1,
# .x=="42292"~ 1,
# .x== "449" ~ 1,
# .x== "78552" ~ 1,
# .x== "7907" ~ 1,
# .x== "9959" ~ 1,
# TRUE ~ 0), .names = "derivdef{.col}"))

# ICD9_0 <- (MSP_clean %>%
# mutate(across(starts_with("diag_cd_"), ~case_when(
# .x="730"~ 1,
# (.x >= 73000) & (.x <= 73009) ~ 1,
# (.x >= 73010) & (.x <= 73019) ~ 1,
# (.x >= 73020) & (.x <= 73029) ~ 1,
# (.x >= 73030) & (.x <= 73039) ~ 1,
# (.x >= 73070) & (.x <= 73078) ~ 1,
# (.x >= 73080) & (.x <= 73089) ~ 1,
# (.x >= 73090) & (.x <= 73099) ~ 1,
# .x="7330"~ 1,
# .x="7301"~ 1,
# .x="7302"~ 1,
# .x="7303"~ 1,
# .x="7307"~ 1,
# .x="7308"~ 1,
# .x="7309"~ 1,
#
# TRUE ~ 0), .names = "derivdef{.col}"))

# ICD9_0 <- (MSP_clean %>%
# mutate(across(starts_with("diag_cd_"), ~case_when(
# (.x >= 0401) & (.x <= 0404) ~ 1,
# (.x >= 45181) & (.x <= 45189) ~ 1,
# .x="0400"~ 1,
# .x="040" ~ 1,
# (.x >= 04081) & (.x <= 04089) ~ 1,
# .x="04041"~ 1,
# .x="04042"~ 1,
# .x="0408" ~ 1,
# .x="324"~ 1,
# .x="3240" ~ 1,
# .x="3241" ~ 1,
# .x="3249"~ 1,
# .x="326"~ 1,
# .x="451" ~ 1,
# (.x >= 45111) & (.x <= 45119) ~ 1,
# .x="4510"~ 1,
# .x="4511" ~ 1,
# .x="4512" ~ 1,
# .x== "4518" ~ 1,
# .x== "4519" ~ 1,
# .x== "567" ~ 1,
# (.x >= 56721) & (x <= 56739) ~ 1,
# (.x >= 5670) & (x <= 5673) ~ 1,
# .x== "5678" ~ 1,
# .x== "56781" ~ 1,
# .x== "56782" ~ 1,
# .x== "56789" ~ 1,
# .x== "5679" ~ 1,
# .x== "5695" ~ 1,
# .x== "572" ~ 1,
#
#
# TRUE ~ 0), .names = "derivde{.col}")))
#
# ICD9_00 <- (MSP_clean %>%
# mutate(across(starts_with("diag_cd_"), ~case_when(
# (.x >= 5720) & (x <= 5728) ~ 1,
# .x== "5901" ~ 1,
# .x== "59010" ~ 1,
# .x== "59011" ~ 1,
# .x== "681" ~ 1,
# .x== "6810" ~ 1,
# .x== "68100" ~ 1,
# .x== "68101" ~ 1,
# .x== "68102" ~ 1,
# .x== "6811" ~ 1,
# .x== "68110" ~ 1,
# .x== "67111" ~ 1,
# .x== "6819" ~ 1,
# .x== "707" ~ 1,
# .x== "7070" ~ 1,
# (.x >= 70700) & (x <= 70709) ~ 1,
# .x== "7071" ~ 1,
# (.x >= 70710) & (x <= 70719) ~ 1,
# .x== "7072" ~ 1,
# (.x >= 70720) & (x <= 70725) ~ 1,
# .x== "7078" ~ 1,
# .x== "7079" ~ 1,
# .x== "7098" ~ 1,
# .x== "7236" ~ 1,
# .x== "72886" ~ 1,
# .x== "7293" ~ 1,
# .x== "72930" ~ 1,
# .x== "72939" ~ 1,
# .x== "7854" ~ 1,
#
# TRUE ~ 0), .names = "derivde{.col}")))
# ICD9_0 <- (MSP_clean %>%
#   mutate(across(starts_with("diag_cd_"), ~case_when(
#     .x=="290"~ 1,
#     .x=="2900"~ 1,
#     .x=="2901"~ 1,
#     (.x >= 29010) & (x <= 29013) ~ 1,
#     .x=="2902"~ 1,
#     .x=="29020"~ 1,
#     .x=="29021"~ 1,
#     .x=="2903"~ 1,
#     .x=="2904"~ 1,
#     .x=="29040"~ 1,
#     .x=="29041"~ 1,
#     .x=="29042"~ 1,
#     .x=="29043"~ 1,
#     .x=="2908"~ 1,
#     .x=="2908~ 1,
#     (.x >= 2930) & (x <= 2949) ~ 1,
#     .x=="310"~ 1,
#     .x=="3100"~ 1,
#     .x=="3101"~ 1,
#     .x=="3102"~ 1,
#     .x=="3108"~ 1,
#     .x=="31081"~ 1,
#     .x=="31089"~ 1,
#     .x=="3109"~ 1,
#     TRUE ~ 0), .names = "deriv{.col}")))

# ICD9_0 <- (MSP_clean %>%
#   mutate(across(starts_with("diag_cd_"), ~case_when(
#     .x=="291"~ 1,
#     (.x >= 2910) & (x <= 2918) ~ 1,
#     (.x >= 29181) & (x <= 29289) ~ 1,
#     .x=="29181"~ 1,
#     .x=="29182"~ 1,
#     .x=="29189"~ 1,
#     .x=="2919"~ 1,
#     .x=="292"~ 1,
#     .x=="2920"~ 1,
#     .x=="2921"~ 1,
#     .x=="29211"~ 1,
#     .x=="29212"~ 1,
#     .x=="2922"~ 1,
#     .x=="2928"~ 1,
#     .x=="29281"~ 1,
#     .x=="29282"~ 1,
# .x=="29283"~1,
# .x=="29284"~1,
# .x=="29289"~1,
# .x=="2929"~1,
# .x=="303"~1,
# .x=="3030"~1,
# .x=="3039"~1,
# .x=="3090"~1,
# .x=="304"~1,
# .x=="3040"~1,
# .x=="30400"~1,
# .x=="30401"~1,
# .x=="3042"~1,
# .x=="30420"~1,
# .x=="30421"~1,
# .x=="30422"~1,
# .x=="30423"~1,
# .x=="3049"~1,
# .x=="3090"~1,
# .x=="30900"~1,
# .x=="30901"~1,
# .x=="30902"~1,
# .x=="30903"~1,
# .x=="30904"~1,
# .x=="30905"~1,
# .x=="30906"~1,
# .x=="30907"~1,
# .x=="30908"~1,
# .x=="30909"~1,
# .x=="3091"~1,
# .x=="30910"~1,
# .x=="30911"~1,
# .x=="30912"~1,
# .x=="30913"~1,
# .x=="30914"~1,
# .x=="30915"~1,
# .x=="30916"~1,
# .x=="30917"~1,
# .x=="30918"~1,
# .x=="30919"~1,
# .x=="3092"~1,
# .x=="30920"~1,
# .x=="30921"~1,
# .x=="30922"~1,
# .x=="30923"~1,
# .x=="30924"~1,
# .x=="30925"~1,
# .x=="30926"~1,
# .x=="30927"~1,
# .x=="30928"~1,
# .x=="30929"~1,
# .x=="3093"~1,
# .x=="30930"~1,
# .x=="30931"~1,
# .x=="30932"~1,
# .x=="30933"~1,
# .x=="30934"~1,
# .x=="30935"~1,
# .x=="30936"~1,
# .x=="30937"~1,
# .x=="30938"~1,
# .x=="30939"~1,
# .x=="3094"~1,
# .x=="30940"~1,
# .x=="30941"~1,
# .x=="30942"~1,
# .x=="30943"~1,
# .x=="30944"~1,
# .x=="30945"~1,
# .x=="30946"~1,
# .x=="3095"~1,
# .x=="30950"~1,
# .x=="3047"~ 1,
# # .x=="30470"~ 1,
# # .x=="30471"~ 1,
# # .x=="30472"~ 1,
# # .x=="30473"~ 1,
# .x=="3048"~ 1,
# # .x=="30480"~ 1,
# # .x=="30481"~ 1,
# # .x=="30482"~ 1,
# # .x=="30483"~ 1,
# .x=="3049"~ 1,
# # .x=="30490"~ 1,
# # .x=="30491"~ 1,
# # .x=="30492"~ 1,
# # .x=="30493"~ 1,
# # .x=="305"~ 1,
# .x=="3050"~ 1,
# # .x=="30500"~ 1,
# # .x=="30501"~ 1,
# # .x=="30502"~ 1,
# # .x=="30503"~ 1,
# .x=="3051"~ 1,
# # .x=="3052"~ 1,
# # .x=="30520"~ 1,
# # .x=="30521"~ 1,
# # .x=="30522"~ 1,
# # .x=="30523"~ 1,
# .x=="3053"~ 1,
# # .x=="30530"~ 1,
# # .x=="30531"~ 1,
# # .x=="30532"~ 1,
# # .x=="30533"~ 1,
# .x=="3054"~ 1,
# # .x=="30540"~ 1,
# # .x=="30541"~ 1,
# # .x=="30542"~ 1,
# # .x=="30543"~ 1,
# .x=="3055"~ 1,
# # .x=="30550"~ 1,
# # .x=="30551"~ 1,
# # .x=="30552"~ 1,
# # .x=="30553"~ 1,
# .x=="3056"~ 1,
# # .x=="30560"~ 1,
# # .x=="30561"~ 1,
# # .x=="30562"~ 1,
# # .x=="30563"~ 1,
# .x=="3057"~ 1,
# # .x=="30570"~ 1,
# # .x=="30571"~ 1,
# # .x=="30572"~ 1,
# TRUE ~ 0), .names = "deriv{.col}"

ICD9_0 <- (MSP_clean %>%
  mutate(across(starts_with("diag_cd_"), ~case_when(
    .x == "2950" ~ 1,
    (.x >= 29500) & (.x <= 29595) ~ 1,
    .x == "29500" ~ 1,
    .x == "29501" ~ 1,
    .x == "29502" ~ 1,
    .x == "29503" ~ 1,
    .x == "29504" ~ 1,
    .x == "29505" ~ 1,
    .x == "2951" ~ 1,
    .x == "29510" ~ 1,
    .x == "29511" ~ 1,
    .x == "29512" ~ 1,
    .x == "29513" ~ 1,
    .x == "29514" ~ 1,
    .x == "29515" ~ 1,
    .x == "2952" ~ 1,
    .x == "29520" ~ 1,
    .x == "29521" ~ 1,
    .x == "29522" ~ 1,
    .x == "29523" ~ 1,
    .x == "29524" ~ 1,
    .x == "29525" ~ 1,
    .x == "2953" ~ 1,
    .x == "29530" ~ 1,
    .x == "29531" ~ 1,
    .x == "29532" ~ 1,
    .x == "29533" ~ 1,
    .x == "29534" ~ 1,
    .x == "29535" ~ 1,
    .x == "2954" ~ 1,
    .x == "29540" ~ 1,})
)
# .x=="29541"~ 1,
# .x=="29542"~ 1,
# .x=="29543"~ 1,
# .x=="29544"~ 1,
# .x=="29545"~ 1,
# .x=="2955"~ 1,
# .x=="29551"~ 1,
# .x=="29552"~ 1,
# .x=="29553"~ 1,
# .x=="29554"~ 1,
# .x=="29555"~ 1,
# .x=="2956"~ 1,
# .x=="29561"~ 1,
# .x=="29562"~ 1,
# .x=="29563"~ 1,
# .x=="29564"~ 1,
# .x=="29565"~ 1,
# .x=="2957"~ 1,
# .x=="29571"~ 1,
# .x=="29572"~ 1,
# .x=="29573"~ 1,
# .x=="29574"~ 1,
# .x=="29575"~ 1,
# .x=="2958"~ 1,
# .x=="29581"~ 1,
# .x=="29582"~ 1,
# .x=="29583"~ 1,
# .x=="29584"~ 1,
# .x=="29585"~ 1,
# .x=="2959"~ 1,
# .x=="29591"~ 1,
# .x=="29592"~ 1,
# .x=="29593"~ 1,
# .x=="29594"~ 1,
# .x=="29595"~ 1,
# .x=="298"~ 1,
# (.x >= 2970) & (.x <= 2989) ~ 1,
# .x=="2980"~ 1,
# .x=="2981"~ 1,
# .x=="2982"~ 1,
# .x=="2983"~ 1,
# .x=="2984"~ 1,
# .x=="2988"~ 1,
# .x=="2989"~ 1,
# .x=="297"~ 1,
# .x=="2970"~ 1,
# .x=="2971"~ 1,
# .x=="2972"~ 1,
# .x=="2973"~ 1,
# .x=="2978"~ 1,
# .x=="2979"~ 1,
# TRUE ~ 0), .names = "derivdef.col")

## ICD9_0 <- (MSP_clean %>%
## mutate(across(starts_with("diag_cd_"), ~case_when(
## .x=="296"~ 1,
## (.x >= 29600) & (.x <= 29699) ~ 1,
## .x=="2960"~ 1,
## .x=="29600"~ 1,
## .x=="29601"~ 1,
## .x=="29602"~ 1,
## .x=="29606"~ 1,
## .x=="2961"~ 1,
## .x=="29610"~ 1,
## .x=="29611"~ 1,
## .x=="29612"~ 1,
## .x=="29613"~ 1,
## .x=="29614"~ 1,
## .x=="29615"~ 1,
## .x=="29616"~ 1,
## .x=="2962"~ 1,
## .x=="29620"~ 1,
## .x=="29621"~ 1,
## .x=="29622"~ 1,
## .x=="29623"~ 1,
## .x=="29624"~ 1,
## .x=="29625"~ 1,
## .x=="29626"~ 1,
## .x=="2963"~ 1,
## .x=="29630"~ 1,
## .x=="29631"~ 1,
## .x=="29632"~ 1,
## .x=="29633"~ 1,
## .x=="29634"~ 1,
## .x=="29635"~ 1,
## .x=="29636"~ 1,
## .x=="2964"~ 1,
## .x=="29640"~ 1,
## .x=="29641"~ 1,
## .x=="29642"~ 1,
## .x=="29643"~ 1,
# ICD9_0 <- (MSP_clean %>%
#     mutate(across(starts_with("diag_cd_"), ~case_when(
#     (.x >= 3000) & (.x <= 3009) ~ 1,
#     # .x=="3000"~ 1,
#     # .x=="3001"~ 1,
#     # .x=="3002"~ 1,
#     # .x=="3003"~ 1,
#     # .x=="3005"~ 1,
#     # .x=="3007"~ 1,
#     # .x=="3008"~ 1,
#     # .x=="3009"~ 1,
#     (.x >= 3060) & (.x <= 3069) ~ 1,
#     # .x=="306"~ 1,
#     # .x=="3060"~ 1,
#     # .x=="3061"~ 1,
#     # .x=="3062"~ 1,
#     # .x=="3063"~ 1,
#     # .x=="3064"~ 1,
#     # .x=="3065"~ 1,
#  (.x >= 30650) & (x <= 30659) ~ 1,
#  .x=="30650"~ 1,
#  .x=="30651"~ 1,
#  .x=="30652"~ 1,
#  .x=="30653"~ 1,
#  .x=="30659"~ 1,
#  .x=="3066"~ 1,
#  .x=="3067"~ 1,
#  .x=="3068"~ 1,
#  .x=="3069"~ 1,
#  .x=="309"~ 1,
#  (.x >= 3090) & (x <= 3099) ~ 1,
#  .x=="3090"~ 1,
#  .x=="3091"~ 1,
#  .x=="3092"~ 1,
#  (.x >= 30921) & (x <= 30929) ~ 1,
#  .x=="30921"~ 1,
#  .x=="30922"~ 1,
#  .x=="30923"~ 1,
#  .x=="30924"~ 1,
#  .x=="30928"~ 1,
#  .x=="30929"~ 1,
#  .x=="3093"~ 1,
#  .x=="3094"~ 1,
#  .x=="3098"~ 1,
#  (.x >= 30981) & (x <= 30989) ~ 1,
#  .x=="30981"~ 1,
#  .x=="30982"~ 1,
#  .x=="30983"~ 1,
#  .x=="30924"~ 1,
#  .x=="30989"~ 1,
#  .x=="3099"~ 1,
#  TRUE ~ 0), .names = "derivde{.col}"

# ICD9_0 <- (MSP_clean %>%
#  mutate(across(starts_with("diag_cd_"), ~case_when(
#    .x=="302.7"~ 1,
#    .x== "307.1" ~ 1,
#    .x== "307.4" ~ 1,
#    .x== "307.5" ~ 1,
#    .x== "316"~ 1,
#    TRUE ~ 0), .names = "derivde{.col}")))
ICD9_0 <- (MSP_clean %>%
  mutate(across(starts_with("diag_cd_"), ~case_when(
    .x == "301" ~ 1,
    .x == "3010" ~ 1,
    .x == "3011" ~ 1,
    .x == "30110" ~ 1,
    .x == "30111" ~ 1,
    .x == "3012" ~ 1,
    .x == "30120" ~ 1,
    .x == "30121" ~ 1,
    .x == "30122" ~ 1,
    .x == "3013" ~ 1,
    .x == "3014" ~ 1,
    .x == "30150" ~ 1,
    .x == "30151" ~ 1,
    .x == "30159" ~ 1,
    .x == "3016" ~ 1,
    .x == "3017" ~ 1,
    .x == "3018" ~ 1,
    .x == "30181" ~ 1,
    .x == "30182" ~ 1,
    .x == "30183" ~ 1,
    .x == "30184" ~ 1,
    .x == "30189" ~ 1,
    .x == "3019" ~ 1,
    .x == "3079" ~ 1,
    .x == "3123" ~ 1,
    TRUE ~ 0), .names = "derivde{.col}")))

ICD9_0 <- (MSP_clean %>%
  mutate(across(starts_with("diag_cd_"), ~case_when(
    .x == "317" ~ 1,
    .x == "318" ~ 1,
    .x == "3180" ~ 1,
    .x == "3181" ~ 1,
    .x == "3182" ~ 1,
    .x == "319" ~ 1,
    TRUE ~ 0), .names = "derivde{.col}")))

ICD9_0 <- (MSP_clean %>%
  mutate(across(starts_with("diag_cd_"), ~case_when(
    .x == "299" ~ 1,
# .x== "2990" ~ 1,
# .x== "29900" ~ 1,
# .x== "29901" ~ 1,
# .x== "2991" ~ 1,
# .x== "29910" ~ 1,
# .x== "29911" ~ 1,
# .x== "2998" ~ 1,
# .x== "29990" ~ 1,
# .x== "29981" ~ 1,
# .x== "2999" ~ 1,
# .x== "29990" ~ 1,
# .x== "29991" ~ 1,
# .x== "315" ~ 1,
# .x== "3150" ~ 1,
# .x== "31500" ~ 1,
# .x== "31501" ~ 1,
# .x== "31502" ~ 1,
# .x== "31509" ~ 1,
# .x== "3151" ~ 1,
# .x== "31532" ~ 1,
# .x== "31534" ~ 1,
# .x== "31535" ~ 1,
# .x== "31539" ~ 1,
# .x== "3154" ~ 1,
# .x== "3155" ~ 1,
# .x== "3158" ~ 1,
# .x== "3159" ~ 1,
#
# TRUE ~ 0), .names = "derivde{.col})"

# ICD9_0 <- (MSP_clean %>%
# mutate(across(starts_with("diag_cd_"), ~case_when(
# (x >= 3070) & (x <= 3076) ~ 1,
# (x >= 31200) & (x <= 31401) ~ 1,
# (x >= 3120) & (x <= 3149) ~ 1,
# # .x== "3070"~ 1,
# # .x== "3072" ~ 1,
# # .x== "3073" ~ 1,
# # .x== "3077" ~ 1,
# # .x== "3076" ~ 1,
# .x== "312" ~ 1,
# # .x== "3120" ~ 1,
# # .x== "31200" ~ 1,
# # .x== "31201" ~ 1,
# # .x== "31202" ~ 1,
# # .x== "31203" ~ 1,
# # .x== "3121" ~ 1,
```bash
# .x== "31210" ~ 1,
# .x== "31211" ~ 1,
# .x== "31212" ~ 1,
# .x== "31213" ~ 1,
# .x== "31220" ~ 1,
# .x== "31221" ~ 1,
# .x== "31222" ~ 1,
# .x== "31223" ~ 1,
# .x== "31224" ~ 1,
# .x== "31230" ~ 1,
# .x== "31231" ~ 1,
# .x== "31232" ~ 1,
# .x== "31233" ~ 1,
# .x== "31234" ~ 1,
# .x== "31235" ~ 1,
# .x== "31239" ~ 1,
# .x== "3124" ~ 1,
# .x== "3128" ~ 1,
# .x== "31281" ~ 1,
# .x== "31282" ~ 1,
# .x== "31289" ~ 1,
# .x== "3129" ~ 1,
# .x== "3130" ~ 1,
# .x== "31300" ~ 1,
# .x== "31301" ~ 1,
# .x== "31302" ~ 1,
# .x== "31321" ~ 1,
# .x== "31322" ~ 1,
# .x== "31323" ~ 1,
# .x== "3133" ~ 1,
# .x== "3138" ~ 1,
# .x== "31381" ~ 1,
# .x== "31382" ~ 1,
# .x== "31383" ~ 1,
# .x== "31389" ~ 1,
# .x== "3138" ~ 1,
# .x== "314" ~ 1,
# .x== "3140" ~ 1,
# .x== "31400" ~ 1,
# .x== "31401" ~ 1,
# .x== "3141" ~ 1,
# .x== "3142" ~ 1,
# .x== "3148" ~ 1,
# .x== "3149" ~ 1,
```

279
# TRUE ~ 0), .names = "derivde{.col}")

ICD9_0 <- (MSP_clean %>%
  mutate(across(starts_with("diag_cd_"), ~case_when(
    .x=="50B"~ 1,
    TRUE ~ 0), .names = "derivde{.col}")))

ICD9_0 <- (MSP_clean %>%
  mutate(across(starts_with("diag_cd_"), ~case_when(
    .x=="291"~ 1,
    .x=="2910"~ 1,
    .x=="2911"~ 1,
    .x=="2912"~ 1,
    .x=="2913"~ 1,
    .x=="2914"~ 1,
    .x=="2915"~ 1,
    .x=="2916"~ 1,
    .x=="2917"~ 1,
    .x=="2918"~ 1,
    .x=="29181"~ 1,
    .x=="29189"~ 1,
    .x=="303"~ 1,
    .x=="3030"~ 1,
    .x=="30301"~ 1,
    .x=="30302"~ 1,
    .x=="30303"~ 1,
    .x=="3039"~ 1,
    .x=="3090"~ 1,
    .x=="3091"~ 1,
    .x=="3092"~ 1,
    .x=="3093"~ 1,
    .x=="305"~ 1,
    .x=="3050"~ 1,
    .x=="30501"~ 1,
    .x=="30502"~ 1,
    .x=="30503"~ 1,
    .x=="3575"~ 1,
    .x=="4255"~ 1,
    .x=="53530"~ 1,
    .x=="53531"~ 1,
    .x=="5710"~ 1,
    .x=="5711"~ 1,
    .x=="5712"~ 1,
    .x=="5713"~ 1,
    .x=="E8600"~ 1,
# TRUE ~ 0), .names = "derivde{.col}"

# ICD9_0 <- (MSP_clean %>%
# mutate(across(starts_with("diag_cd_"), ~case_when(  
#   .x=="0400"~ 1,
#   .x=="3040"~ 1,
#   (.x >= 30401) & (.x <= 30473) ~ 1,
#   (.x >= 30550) & (.x <= 30553) ~ 1,
#   (.x >= 96500) & (.x <= 96509) ~ 1,
#   .x=="30401"~ 1,
#   .x=="30402"~ 1,
#   .x=="30403"~ 1,
#   .x=="30470"~ 1,
#   .x=="30471"~ 1,
#   .x=="30472"~ 1,
#   .x=="30473"~ 1,
#   .x=="30550"~ 1,
#   .x=="30551"~ 1,
#   .x=="30552"~ 1,
#   .x=="30553"~ 1,
#   .x=="96500"~ 1,
#   .x=="96501"~ 1,
#   .x=="96502"~ 1,
#   .x=="96509"~ 1,
#   .x=="9701"~ 1,
#   .x=="E8500"~ 1,
#   .x=="E8501"~ 1,
#   .x=="E8502"~ 1,
#   .x=="E9350"~ 1,
#   .x=="E9351"~ 1,
#   .x=="E9352"~ 1,
#   .x=="E9401"~ 1,
#   # TRUE ~ 0), .names = "derivde{.col}"))
# ICD9_0 <- (MSP_clean %>%
#   mutate(across(starts_with("diag_cd_"), ~case_when(
#     .x=="3043"~ 1,
#     .x=="30430"~ 1,
#     .x=="30431"~ 1,
#     .x=="30432"~ 1,
#     .x=="30433"~ 1,
#     .x=="3052"~ 1,
#     .x=="30520"~ 1,
#     .x=="30521"~ 1,
#     .x=="30522"~ 1,
#     .x=="30523"~ 1,
#     TRUE ~ 0), .names = "derivde{.col}")))

# ICD9_0 <- (MSP_clean %>%
#   mutate(across(starts_with("diag_cd_"), ~case_when(
#     .x=="3041"~ 1,
#     .x=="30410"~ 1,
#     .x=="30411"~ 1,
#     .x=="30412"~ 1,
#     .x=="30413"~ 1,
#     .x=="3054"~ 1,
#     .x=="30540"~ 1,
#     .x=="30541"~ 1,
#     .x=="30542"~ 1,
#     .x=="30543"~ 1,
#     TRUE ~ 0), .names = "derivde{.col}")))

# ICD9_0 <- (MSP_clean %>%
#   mutate(across(starts_with("diag_cd_"), ~case_when(
#     .x=="3042"~ 1,
#     .x=="30420"~ 1,
#     .x=="30421"~ 1,
# .x="30422"~ 1,
# .x="30423"~ 1,
# .x="3056"~ 1,
# .x="30560"~ 1,
# .x="30561"~ 1,
# .x="30562"~ 1,
# .x="30563"~ 1,
# .x="9685"~ 1,
# .x="E9385"~ 1,
#
# # TRUE ~ 0), .names = "derivde{.col}")))

# ICD9_0 <- (MSP_clean %>%
# mutate(across(starts_with("diag_cd _"), ~case_when(
#   .x="3044"~ 1,
#   .x="30440"~ 1,
#   .x="30441"~ 1,
#   .x="30442"~ 1,
#   .x="30443"~ 1,
#   .x="3057"~ 1,
#   .x="30570"~ 1,
#   .x="30571"~ 1,
#   .x="30572"~ 1,
#   .x="30573"~ 1,
#   TRUE ~ 0), .names = "derivde{.col}")))

# ICD9_0 <- (MSP_clean %>%
# mutate(across(starts_with("diag_cd _"), ~case_when(
#   .x="3045"~ 1,
#   .x="30450"~ 1,
#   .x="30451"~ 1,
#   .x="30452"~ 1,
#   .x="30453"~ 1,
#   .x="3053"~ 1,
#   .x="30530"~ 1,
#   .x="30531"~ 1,
#   .x="30532"~ 1,
#   .x="30533"~ 1,
#   .x="9696"~ 1,
# .x=="E9396"~ 1,
# .x=="E8541"~ 1,
#
#
# TRUE ~ 0), .names = "derivde{.col}"
#
# ICD9_0 <- (MSP_clean %>%
# mutate(across(starts_with("diag_cd_")), ~case_when(
# .x=="3051"~ 1,
# .x=="V1582"~ 1,
# TRUE ~ 0), .names = "derivde{.col}"
# ))
# ICD9_0 <- (MSP_clean %>%
# mutate(across(starts_with("diag_cd_")), ~case_when(
# (.x >= 64830) & (.x <= 64834) ~ 1,
# (.x >= 30461) & (.x <= 30463) ~ 1,
# (.x >= 30591) & (.x <= 30593) ~ 1,
# TRUE ~ 0), .names = "derivde{.col}"
# ))
# ICD9_0 <- (MSP_clean %>%
# mutate(across(starts_with("diag_cd_")), ~case_when(
# (.x >= 30491) & (.x <= 30493) ~ 1,
# (.x >= 7960) & (.x <= 7960) ~ 1,
# TRUE ~ 0), .names = "derivde{.col}"
# ))
# ICD9_00 <- (MSP_clean %>%
# mutate(across(starts_with("diag_cd_")), ~case_when(
# .x=="292"~ 1,
# .x=="6483"~ 1,
# .x=="7960"~ 1,
# .x=="9621"~ 1,
# TRUE ~ 0), .names = "derivde{.col}"
# ))
# .x=="9658" ~ 1,
# .x=="9663" ~ 1,
# .x=="9664" ~ 1,
# .x=="9670" ~ 1,
# .x=="9684" ~ 1,
# .x=="9685" ~ 1,
# (.x >= 9696) & (.x <= 9699) ~ 1,
# .x=="970" ~ 1,
#
# TRUE ~ 0), .names = "derivde{.col}"))

# icd9_1 <- ICD9_0 %>%
# group_by(moh_study_id) %>%
# summarize(derivdediag_cd_1 = sum(derivdediag_cd_1),
# derivdediag_cd_2 = sum(derivdediag_cd_2),
# derivdediag_cd_3 = sum(derivdediag_cd_3))
#
# icd9_11 <- ICD9_00 %>%
# group_by(moh_study_id) %>%
# summarize(derivdediag_cd_1 = sum(derivdediag_cd_1),
# derivdediag_cd_2 = sum(derivdediag_cd_2),
# derivdediag_cd_3 = sum(derivdediag_cd_3))
#
# icd9_2 <- icd9_1 %>%
# mutate(Total = select(.,-moh_study_id) %>%
# rowSums(na.rm = TRUE))
#
# icd9_22 <- icd9_11 %>%
# mutate(Total1 = select(.,-moh_study_id) %>%
# rowSums(na.rm = TRUE))
#
# icd9_222 <- merge(x=icd9_2, y= icd9_22, by="moh_study_id", all=TRUE)
#
# icd9_2222 <- icd9_222 %>%
# mutate(Total3 = Total + Total1)
# icd9_continous <- icd9_2222 %>%
# mutate(Polysubstance = Total3) %>%
# select(moh_study_id, Polysubstance)
#
# icd9_3 <- icd9_2222 %>%
# mutate(Polysubstance = if_else(Total3 >= 1, 1, 0))
#
# icd9_binary <- icd9_3 %>%
# select(moh_study_id, Polysubstance)
#
# table(icd9_continous$Polysubstance)
```r
# table(icd9_binary$Polysubstance)
#
# write.csv(icd9_continous, "Icd9_Polysubstance_C.csv", row.names = FALSE)
# write.csv(icd9_binary, "Icd9_Polysubstance_B.csv", row.names = FALSE)
#
# icd9_1 <- ICD9_0 %>%
#   group_by(moh_study_id) %>%
#   summarize(derivdediag_cd_1 = sum(derivdediag_cd_1),
#             derivdediag_cd_2 = sum(derivdediag_cd_2),
#             derivdediag_cd_3 = sum(derivdediag_cd_3))
#
# icd9_2 <- icd9_1 %>%
#   mutate(Total = select(., moh_study_id) %>%
#           rowSums(na.rm = TRUE))
#
# icd9_continious <- icd9_2 %>%
#   mutate(Polysubstance = Total) %>%
#   select(moh_study_id, Polysubstance)
#
# icd9_3 <- icd9_2 %>%
#   mutate(Polysubstance = if_else(Total >= 1, 1, 0))
#
# icd9_binary <- icd9_3 %>%
#   select(moh_study_id, Polysubstance)
#
# table(icd9_continious$Polysubstance)
# table(icd9_binary$Polysubstance)
#
# write.csv(icd9_continious, "Icd9_Polysubstance_C.csv", row.names = FALSE)
# write.csv(icd9_binary, "Icd9_Polysubstance_B.csv", row.names = FALSE)
#
# table(Icd9_cocaine_B$cocaine)
# table(Icd9_opiates_B$opiates)
```

Icd Code 10

```r
library(tidyverse)
```
library(dplyr)
library(gapminder)
library(caret)

## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
## lift

library(xgboost)

##
## Attaching package: 'xgboost'
##
## The following object is masked from 'package:dplyr':
##
## slice

library(forcats)
library(mice)

##
## Attaching package: 'mice'
##
## The following object is masked from 'package:stats':
##
## filter
##
## The following objects are masked from 'package:base':
##
## cbind, rbind

library(vctrs)

##
## Attaching package: 'vctrs'
##
## The following object is masked from 'package:dplyr':
##
## data_frame
##
## The following object is masked from 'package:tibble':
##
## data_frame

###Data descriptive
vw_dad <- read_csv("U:/Andy's Thesis/Raw_data_used/vw_dad.csv", show_col_types = FALSE)

## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
## dat <- vroom(...)
## problems(dat)

glimpse(vw_dad)

## Rows: 1,663,940
## Columns: 141

###Template

```r
# ICD10_0 <- (vw_dad %>%
#   mutate(across(starts_with("diagx_"), ~case_when(
#     .x=="E8500"~ 1,
#     .x== "3040" ~ 1,
#     .x== "3047" ~ 1,
#     .x=="3055"~ 1,
#     .x== "9650" ~ 1,
#     TRUE ~ 0), .names = "derivde{.col}")))

# ICD10_00 <- (vw_dad %>%
#   mutate(across(starts_with("diagx_"), ~case_when(
#     (.x >= 301) & (x <= 301.9) ~ 1,
#     TRUE ~ 0), .names = "derivde{.col}")))

# ICD10_0 <- (vw_dad %>%
#              mutate(across(starts_with("diagx_"), ~case_when(
#                grepl("F10", .) ~ 1,
#                TRUE ~ 0), .names = "derived{.col}")))

###start extracting ICD10

# ICD10_0 <- (vw_dad %>%
#               mutate(across(starts_with("diagx_"), ~case_when(
#                 grepl("B376", .) ~ 1,
#                 grepl("I33", .) ~ 1,
#                 grepl("I34", .) ~ 1,
#                 grepl("I35", .) ~ 1,
#                 grepl("I36", .) ~ 1,
#                 grepl("I37", .) ~ 1,
#                 grepl("I38", .) ~ 1,
#                 grepl("I39", .) ~ 1,
#                 TRUE ~ 0), .names = "derived{.col}")))

# ICD10_0 <- (vw_dad %>%
#               mutate(across(starts_with("diagx_"), ~case_when(
#                 grepl("A40", .) ~ 1,
#                 grepl("A41", .) ~ 1,
```

288
```r
# grepl("I269", .) ~ 1,
# grepl("I400", .) ~ 1,
# grepl("R572", .) ~ 1,
# grepl("R651", .) ~ 1,
# grepl("R659", .) ~ 1,
# TRUE ~ 0), .names = "derived{.col}"))

# ICD10_0 <- (vw_dad %>%
# mutate(across(starts_with("diagx_"), ~case_when(
# grepl("M86", .) ~ 1,
# grepl("M899", .) ~ 1,
# TRUE ~ 0), .names = "derived{.col}"))

# ICD10_0 <- (vw_dad %>%
# mutate(across(starts_with("diagx_"), ~case_when(
# grepl("I80", .) ~ 1,
# grepl("L97", .) ~ 1,
# grepl("L988", .) ~ 1,
# grepl("M793", .) ~ 1,
# grepl("A480", .) ~ 1,
# grepl("G06", .) ~ 1,
# grepl("G09", .) ~ 1,
# grepl("K630", .) ~ 1,
# grepl("K650", .) ~ 1,
# grepl("K750", .) ~ 1,
# grepl("L02", .) ~ 1,
# grepl("L03", .) ~ 1,
# grepl("M5402", .) ~ 1,
# grepl("M726", .) ~ 1,
# grepl("N10", .) ~ 1,
# grepl("R02", .) ~ 1,
# TRUE ~ 0), .names = "derived{.col}"))

# ICD10_0 <- (vw_dad %>%
# mutate(across(starts_with("diagx_"), ~case_when(
# grepl("F00", .) ~ 1,
# grepl("F01", .) ~ 1,
# grepl("F02", .) ~ 1,
# grepl("F03", .) ~ 1,
# grepl("F04", .) ~ 1,
# grepl("F05", .) ~ 1,
# grepl("F06", .) ~ 1,
# grepl("F07", .) ~ 1,
# grepl("F08", .) ~ 1,
# grepl("F09", .) ~ 1,
# grepl("G30", .) ~ 1,
```

# TRUE ~ 0), .names = "derived{.col}"))

# ICD10_0 <- vw_dad %>%
# mutate(across(starts_with("diagx "), ~case_when(
# grepl("F10", .) ~ 1,
# grepl("F11", .) ~ 1,
# grepl("F12", .) ~ 1,
# grepl("F13", .) ~ 1,
# grepl("F14", .) ~ 1,
# grepl("F15", .) ~ 1,
# grepl("F16", .) ~ 1,
# grepl("F17", .) ~ 1,
# grepl("F18", .) ~ 1,
# grepl("F19", .) ~ 1,
# grepl("F55", .) ~ 1,
# grepl("T40", .) ~ 1,
#
# TRUE ~ 0), .names = "derived{.col}"))

# ICD10_0 <- vw_dad %>%
# mutate(across(starts_with("diagx "), ~case_when(
# grepl("F20", .) ~ 1,
# grepl("F21", .) ~ 1,
# grepl("F22", .) ~ 1,
# grepl("F23", .) ~ 1,
# grepl("F24", .) ~ 1,
# grepl("F25", .) ~ 1,
# grepl("F26", .) ~ 1,
# grepl("F27", .) ~ 1,
# grepl("F28", .) ~ 1,
# grepl("F29", .) ~ 1,
#
# TRUE ~ 0), .names = "derived{.col}"))

# ICD10_0 <- vw_dad %>%
# mutate(across(starts_with("diagx "), ~case_when(
# grepl("F30", .) ~ 1,
# grepl("F31", .) ~ 1,
# grepl("F32", .) ~ 1,
# grepl("F33", .) ~ 1,
# grepl("F34", .) ~ 1,
# grepl("F35", .) ~ 1,
# grepl("F36", .) ~ 1,
# grepl("F37", .) ~ 1,
# grepl("F38", .) ~ 1,
# grepl("F39", .) ~ 1,
# ICD10_0 <- (vw_dad %>%
#    mutate(across(starts_with("diagx_{"}, ~case_when(
#      grepl("F40", .) ~ 1,
#      grepl("F41", .) ~ 1,
#      grepl("F42", .) ~ 1,
#      grepl("F43", .) ~ 1,
#      grepl("F44", .) ~ 1,
#      grepl("F45", .) ~ 1,
#      grepl("F46", .) ~ 1,
#      grepl("F47", .) ~ 1,
#      grepl("F48", .) ~ 1,
#      TRUE ~ 0), .names = "derived{.col}")))

# ICD10_0 <- (vw_dad %>%
#    mutate(across(starts_with("diagx_{"), ~case_when(
#      grepl("F50", .) ~ 1,
#      grepl("F51", .) ~ 1,
#      grepl("F52", .) ~ 1,
#      grepl("F53", .) ~ 1,
#      grepl("F54", .) ~ 1,
#      grepl("F55", .) ~ 1,
#      grepl("F56", .) ~ 1,
#      grepl("F57", .) ~ 1,
#      grepl("F58", .) ~ 1,
#      grepl("F59", .) ~ 1,
#      TRUE ~ 0), .names = "derived{.col}")))

# ICD10_0 <- (vw_dad %>%
#    mutate(across(starts_with("diagx_{"), ~case_when(
#      grepl("F60", .) ~ 1,
#      grepl("F61", .) ~ 1,
#      grepl("F62", .) ~ 1,
#      grepl("F63", .) ~ 1,
#      grepl("F64", .) ~ 1,
#      grepl("F65", .) ~ 1,
#      grepl("F66", .) ~ 1,
#      grepl("F67", .) ~ 1,
#      grepl("F68", .) ~ 1,
#      grepl("F69", .) ~ 1,
#      TRUE ~ 0), .names = "derived{.col}")))
```r
# ICD10_0 <- (vw_dad %>%
#   mutate(across(starts_with("diagx_"), ~case_when(
#     grepl("F70", .) ~ 1,
#     grepl("F71", .) ~ 1,
#     grepl("F72", .) ~ 1,
#     grepl("F73", .) ~ 1,
#     grepl("F74", .) ~ 1,
#     grepl("F75", .) ~ 1,
#     grepl("F76", .) ~ 1,
#     grepl("F77", .) ~ 1,
#     grepl("F78", .) ~ 1,
#     grepl("F79", .) ~ 1,
#     TRUE ~ 0), .names = "derived{.col}")))
# ICD10_0 <- (vw_dad %>%
#   mutate(across(starts_with("diagx_"), ~case_when(
#     grepl("F80", .) ~ 1,
#     grepl("F81", .) ~ 1,
#     grepl("F82", .) ~ 1,
#     grepl("F83", .) ~ 1,
#     grepl("F84", .) ~ 1,
#     grepl("F85", .) ~ 1,
#     grepl("F86", .) ~ 1,
#     grepl("F87", .) ~ 1,
#     grepl("F88", .) ~ 1,
#     grepl("F89", .) ~ 1,
#     TRUE ~ 0), .names = "derived{.col}")))
# ICD10_0 <- (vw_dad %>%
#   mutate(across(starts_with("diagx_"), ~case_when(
#     grepl("F90", .) ~ 1,
#     grepl("F91", .) ~ 1,
#     grepl("F92", .) ~ 1,
#     grepl("F93", .) ~ 1,
#     grepl("F94", .) ~ 1,
#     grepl("F95", .) ~ 1,
#     grepl("F96", .) ~ 1,
#     grepl("F97", .) ~ 1,
#     grepl("F98", .) ~ 1,
#     grepl("F99", .) ~ 1,
#     TRUE ~ 0), .names = "derived{.col}")))
# ICD10_0 <- (vw_dad %>%
#   mutate(across(starts_with("diagx_"), ~case_when(
#     grepl("F99", .) ~ 1,
```
ICD10_0 <- (vw_dad %>%
  mutate(across(starts_with("diagx_"), ~case_when(
    grepl("F10", .) ~ 1,
    grepl("T400", .) ~ 1,
    grepl("T401", .) ~ 1,
    grepl("T402", .) ~ 1,
    grepl("T403", .) ~ 1,
    grepl("Z79891", .) ~ 1,
    TRUE ~ 0), .names = "derived{.col}")))

ICD10_0 <- (vw_dad %>%
  mutate(across(starts_with("diagx_"), ~case_when(
    grepl("F11", .) ~ 1,
    TRUE ~ 0), .names = "derived{.col}")))

ICD10_0 <- (vw_dad %>%
  mutate(across(starts_with("diagx_"), ~case_when(
    grepl("F12", .) ~ 1,
    TRUE ~ 0), .names = "derived{.col}")))

ICD10_0 <- (vw_dad %>%
  mutate(across(starts_with("diagx_"), ~case_when(
    grepl("F13", .) ~ 1,
    TRUE ~ 0), .names = "derived{.col}")))

ICD10_0 <- (vw_dad %>%
  mutate(across(starts_with("diagx_"), ~case_when(
    grepl("F14", .) ~ 1,
    TRUE ~ 0), .names = "derived{.col}")))

ICD10_0 <- (vw_dad %>%
  mutate(across(starts_with("diagx_"), ~case_when(
    grepl("F15", .) ~ 1,
    TRUE ~ 0), .names = "derived{.col}")))

ICD10_0 <- (vw_dad %>%
  mutate(across(starts_with("diagx_"), ~case_when(
    grepl("F16", .) ~ 1,
```r
# grepl("T408", ) ~ 1,
# TRUE ~ 0), .names = "derived{.col}")

# ICD10_0 <- (vw_dad %>%
#   mutate(across(starts_with("diagx_"), ~case_when(
#     grepl("F17", ) ~ 1,
#     TRUE ~ 0), .names = "derived{.col}")

# ICD10_0 <- (vw_dad %>%
#   mutate(across(starts_with("diagx_"), ~case_when(
#     grepl("F18", ) ~ 1,
#     TRUE ~ 0), .names = "derived{.col}")

# ICD10_0 <- (vw_dad %>%
#   mutate(across(starts_with("diagx_"), ~case_when(
#     grepl("V6542", ) ~ 1,
#     grepl("F19", ) ~ 1,
#     grepl("Z715", ) ~ 1,
#     grepl("Z503", ) ~ 1,
#     grepl("T42", ) ~ 1,
#     grepl("T387", ) ~ 1,
#     grepl("T408", ) ~ 1,
#     grepl("T409", ) ~ 1,
#     grepl("T412", ) ~ 1,
#     grepl("T436", ) ~ 1,
#     grepl("T437", ) ~ 1,
#     grepl("T438", ) ~ 1,
#     grepl("T439", ) ~ 1,
#     grepl("T507", ) ~ 1,
#     TRUE ~ 0), .names = "derived{.col}")

# ICD10_1 <- (ICD10_0 %>%
#   select(moh_study_id, starts_with("derived"))
# # ICD10_2 <- (ICD10_1 %>%
# #   group_by(moh_study_id) %>%
# #   summarize(deriveddiagx_1 = sum(deriveddiagx_1),
# # deriveddiagx_2 = sum(deriveddiagx_2),
# # deriveddiagx_3 = sum(deriveddiagx_3),
# # deriveddiagx_4 = sum(deriveddiagx_4),
# # deriveddiagx_5 = sum(deriveddiagx_5),
# # deriveddiagx_6 = sum(deriveddiagx_6),
# # deriveddiagx_7 = sum(deriveddiagx_7),
# # deriveddiagx_8 = sum(deriveddiagx_8),
# # deriveddiagx_9 = sum(deriveddiagx_9),
# # deriveddiagx_10 = sum(deriveddiagx_10),
# # deriveddiagx_11 = sum(deriveddiagx_11),
```
# deriveddiagx_12 = sum(deriveddiagx_12),
# deriveddiagx_13 = sum(deriveddiagx_13),
# deriveddiagx_14 = sum(deriveddiagx_14),
# deriveddiagx_15 = sum(deriveddiagx_15),
# deriveddiagx_16 = sum(deriveddiagx_16),
# deriveddiagx_17 = sum(deriveddiagx_17),
# deriveddiagx_18 = sum(deriveddiagx_18),
# deriveddiagx_19 = sum(deriveddiagx_19),
# deriveddiagx_20 = sum(deriveddiagx_20),
# deriveddiagx_21 = sum(deriveddiagx_21),
# deriveddiagx_22 = sum(deriveddiagx_22),
# deriveddiagx_23 = sum(deriveddiagx_23),
# deriveddiagx_24 = sum(deriveddiagx_24),
# deriveddiagx_25 = sum(deriveddiagx_25))
#
# ICD10_3 <- ICD10_2 %>%
#   mutate(Total = select(.,-moh_study_id) %>% rowSums(na.rm = TRUE))
#
# icd10_continuous <- ICD10_3 %>%
#   mutate(Tobaccouse = Total) %>
#   select(moh_study_id, Tobaccouse)
#
# ICD10_4 <- ICD10_3 %>%
#   mutate(Tobaccouse = if_else( Total >= 1, 1, 0))
#
# icd10_binary <- ICD10_4 %>%
#   select(moh_study_id, Tobaccouse)
#
# table(icd10_continuous$Tobaccouse)
# table(icd10_binary$Tobaccouse)
#
# write.csv(icd10_continuous, "Icd10_Tobaccouse_C.csv", row.names = FALSE)
# write.csv(icd10_binary, "Icd10_Tobaccouse_B.csv", row.names = FALSE)

Pharmanet

# Pharmanetspss <- read_csv("U:/Andy's Thesis/Raw_data_used/Pharmanetspss.csv",
#  , show_col_types = FALSE)
# glimpse(Pharmanetspss)

### OST/OAT

# dinpin <- (Pharmanetspss %>%
#   mutate(across(starts_with("din_pin"), ~case_when(
#     .x="2242963"~ 1,
#     .x="2242964"~ 1,
#     .x="2295695"~ 1,
#     .x="2295709"~ 1,
#     .x="2408104"~ 1,
# 
# .x="2424851"~ 1,
# .x="2424878"~ 1,
# .x="999776"~ 1,
# .x="655627"~ 1,
# .x="781460"~ 1,
# .x="781479"~ 1,
# .x="2408090"~ 1,
# (.x >= 999792) & (.x <= 999793) ~ 1,
# (.x >= 66999990) & (.x <= 66999993) ~ 1,
# (.x >= 66999997) & (.x <= 66999999) ~ 1,
# (.x >= 67000000) & (.x <= 67000004) ~ 1,
# (.x >= 22123346) & (.x <= 22123349) ~ 1,
# TRUE ~ 0, .names = "derivde{.col}")))
# dinpin_1 <- dinpin %>%
# rename(ost = derivdedin_pin) %>
# select(moh_study_id, oat, ost)
# write.csv(dinpin_1, "Pharmanetdinpin.csv", row.names = FALSE)

Combining Icd files

# Tobaccouse1 <- left_join(Icd9_Tobaccouse_B, Icd10_Tobaccouse_B, by="moh_study_id") %>
# mutate_at(c(3), ~replace_na(.,0))
# Tobaccouse <- Tobaccouse1 %>
# mutate(Tobaccouse = Tobaccouse.x + Tobaccouse.y) %>
# mutate(Tobaccouse = if_else(Tobaccouse >= 1, 1, 0)) %>
# select(moh_study_id, Tobaccouse)
# write.csv(Tobaccouse, "Tobaccouse.csv", row.names = FALSE)
# table(Tobaccouse$Tobaccouse)

# Mooddisorders1 <- left_join(lcd9_Mooddisorders_B, Icd10_Mooddisorders_B, by="moh_study_id") %>
# mutate_at(c(3), ~replace_na(.,0))
# Mooddisorders <- Mooddisorders1 %>
# mutate(Mooddisorders = Mooddisorders.x + Mooddisorders.y) %>
# mutate(Mooddisorders = if_else(Mooddisorders >= 1, 1, 0)) %>
# select(moh_study_id, Mooddisorders)
# write.csv(Mooddisorders, "Mooddisorders.csv", row.names = FALSE)
# table(Mooddisorders$Mooddisorders)
# Tissueinfection1 <- left_join(Icd9_Tissueinfection_B, Icd10_Tissueinfection_B, by="moh_study_id") %>%
#  mutate_at(c(3), ~replace_na(.,0))
#
# Tissueinfection <- Tissueinfection1 %>%
#  mutate(Tissueinfection = Tissueinfection.x + Tissueinfection.y) %>%
#  mutate(Tissueinfection = if_else(Tissueinfection >= 1, 1, 0)) %>%
#  select(moh_study_id, Tissueinfection)
#
# write.csv(Tissueinfection, "Tissueinfection.csv", row.names = FALSE)
#
# table(Tissueinfection$Tissueinfection)
#
# substancerelateddisorders <- left_join(Icd9_substancerelateddisorders_B, Icd10_substancerelateddisorders_B, by="moh_study_id") %>%
#  mutate_at(c(3), ~replace_na(.,0))
#
# substancerelateddisorders <- substancerelateddisorders %>%
#  mutate(substancerelateddisorders = substancerelateddisorders.x + substancerelateddisorders.y) %>%
#  mutate(substancerelateddisorders = if_else(substancerelateddisorders >= 1, 1, 0)) %>%
#  select(moh_study_id, substancerelateddisorders)
#
# write.csv(substancerelateddisorders, "substancerelateddisorders.csv", row.names = FALSE)
#
# table(substancerelateddisorders$substancerelateddisorders)
#
# ICDderivedvariables$results = ifelse(ICDderivedvariables$substancerelateddisorders.x > ICDderivedvariables$substancerelateddisorders.y, "column1",
#  ifelse(ICDderivedvariables$substancerelateddisorders.x < ICDderivedvariables$substancerelateddisorders.y, 'Column2', 'None'))
#
# table(ICDderivedvariables$results)
#
# Stimulantuse1 <- left_join(Icd9_Stimulantuse_B, Icd10_Stimulantuse_B, by="moh_study_id") %>%
#  mutate_at(c(3), ~replace_na(.,0))
#
# Stimulantuse <- Stimulantuse1 %>%
#  mutate(Stimulantuse = Stimulantuse.x + Stimulantuse.y) %>%
#  mutate(Stimulantuse = if_else(Stimulantuse >= 1, 1, 0)) %>%
#  select(moh_study_id, Stimulantuse)
#
# write.csv(Stimulantuse, "Stimulantuse.csv", row.names = FALSE)
#
# table(Stimulantuse$Stimulantuse)
# length(unique(Stimulantuse$moh_study_id)) == nrow(Stimulantuse)

# Sepsis1 <- left_join(Icd9_Sepsis_B, Icd10_Sepsis_B, by="moh_study_id") %>%
#   mutate_at(c(3), ~replace_na(.,0))
#
# Sepsis <- Sepsis1 %>%
#   mutate(Sepsis = Sepsis.x + Sepsis.y) %>%
#   mutate(Sepsis = if_else(Sepsis >= 1, 1, 0)) %>%
#   select(moh_study_id, Sepsis)
#
# write.csv(Sepsis, "Sepsis.csv", row.names = FALSE)
#
# length(unique(Sepsis$moh_study_id)) == nrow(Sepsis)
#
# Sedativeandhypnoticuse1 <- left_join(Icd9_Sedativeandhypnoticuse_B, Icd10_Sedativeandhypnoticuse_B, by="moh_study_id") %>%
#   mutate_at(c(3), ~replace_na(.,0))
#
# Sedativeandhypnoticuse <- Sedativeandhypnoticuse1 %>%
#   mutate(Sedativeandhypnoticuse = Sedativeandhypnoticuse.x + Sedativeandhypnoticuse.y) %>%
#   mutate(Sedativeandhypnoticuse = if_else(Sedativeandhypnoticuse >= 1, 1, 0)) %>%
#   select(moh_study_id, Sedativeandhypnoticuse)
#
# write.csv(Sedativeandhypnoticuse, "Sedativeandhypnoticuse.csv", row.names = FALSE)
#
# length(unique(Sedativeandhypnoticuse$moh_study_id)) == nrow(Sedativeandhypnoticuse)
#
# psychoticdisorders1 <- left_join(Icd9_psychoticdisorders_B, Icd10_psychoticdisorders_B, by="moh_study_id") %>%
#   mutate_at(c(3), ~replace_na(.,0))
#
# psychoticdisorders <- psychoticdisorders1 %>%
#   mutate(psychoticdisorders = psychoticdisorders.x + psychoticdisorders.y) %>%
#   mutate(psychoticdisorders = if_else(psychoticdisorders >= 1, 1, 0)) %>%
#   select(moh_study_id, psychoticdisorders)
#
# write.csv(psychoticdisorders, "psychoticdisorders.csv", row.names = FALSE)
#
# length(unique(psychoticdisorders$moh_study_id)) == nrow(psychoticdisorders)
#
# Polysubstance1 <- left_join(Icd9_Polysubstance_B, Icd10_Polysubstance_B, by="moh_study_id") %>%
#   mutate_at(c(3), ~replace_na(.,0))
#
# Polysubstance <- Polysubstance1 %>%
# mutate(Polysubstance = Polysubstance.x + Polysubstance.y) %>%
# mutate(Polysubstance = if_else(Polysubstance >= 1, 1, 0)) %>%
# select(moh_study_id, Polysubstance)
#
# write.csv(Polysubstance, "Polysubstance.csv", row.names = FALSE)
#
# table(Polysubstance$Polysubstance)
#
# length(unique(Polysubstance$moh_study_id)) == nrow(Polysubstance)
#
# Personalitydisorders1 <- left_join(Icd9_Personalitydisorders_B, Icd10_Personalitydisorders_B, by="moh_study_id") %>%
# mutate_at(c(3), ~replace_na(.,0))
#
# Personalitydisorders < Personalitydisorders1 %>%
# mutate(Personalitydisorders = Personalitydisorders.x + Personalitydisorders.y) %>%
# mutate(Personalitydisorders = if_else(Personalitydisorders >= 1, 1, 0)) %>%
# select(moh_study_id, Personalitydisorders)
#
# write.csv(Personalitydisorders, "Personalitydisorders.csv", row.names = FALSE)
#
# table(Personalitydisorders$Personalitydisorders)
#
# length(unique(Personalitydisorders$moh_study_id)) == nrow(Personalitydisorders)
#
# Otherpsychoactivedruguse1 <- left_join(Icd9_Otherpsychoactivedruguse_B, Icd10_Otherpsychoactivedruguse_B, by="moh_study_id") %>%
# mutate_at(c(3), ~replace_na(.,0))
#
# Otherpsychoactivedruguse < Otherpsychoactivedruguse1 %>%
# mutate(Otherpsychoactivedruguse = Otherpsychoactivedruguse.x + Otherpsychoactivedruguse.y) %>%
# mutate(Otherpsychoactivedruguse = if_else(Otherpsychoactivedruguse >= 1, 1, 0)) %>%
# select(moh_study_id, Otherpsychoactivedruguse)
#
# write.csv(Otherpsychoactivedruguse, "Otherpsychoactivedruguse.csv", row.names = FALSE)
#
# table(Otherpsychoactivedruguse$Otherpsychoactivedruguse)
#
# length(unique(Otherpsychoactivedruguse$moh_study_id)) == nrow(Otherpsychoactivedruguse)
#
# Osteomyelitis1 <- left_join(Icd9_Osteomyelitis_B, Icd10_Osteomyelitis_B, by="moh_study_id") %>%
# mutate_at(c(3), ~replace_na(.,0))
#
# Osteomyelitis < Osteomyelitis1 %>%
# mutate(Osteomyelitis = Osteomyelitis.x + Osteomyelitis.y) %>%
# mutate(Osteomyelitis = if_else(Osteomyelitis >= 1, 1, 0)) %>%
# select(moh_study_id, Osteomyelitis)
#
# write.csv(Osteomyelitis, "Osteomyelitis.csv", row.names = FALSE)
# table(Osteomyelitis$Osteomyelitis)
#
# length(unique(Osteomyelitis$moh_study_id)) == nrow(Osteomyelitis)

# Opioiduse1 <- left_join(Icd9_Opioiduse_B, Icd10_Opioiduse_B, by="moh_study_id") %>%
#   mutate_at(c(3), ~replace_na(.,0))
#
# Opioiduse <- Opioiduse1 %>%
#   mutate(Opioiduse = Opioiduse.x + Opioiduse.y) %>%
#   mutate(Opioiduse = if_else(Opioiduse >= 1, 1, 0)) %>%
#   select(moh_study_id, Opioiduse)
#
# write.csv(Opioiduse, "Opioiduse.csv", row.names = FALSE)
#
# table(Opioiduse$Opioiduse)
#
# length(unique(Opioiduse$moh_study_id)) == nrow(Opioiduse)

# Neuroticrelateddisorders1 <- left_join(Icd9_Neuroticrelateddisorders_B, Icd10_Neuroticrelateddisorders_B, by="moh_study_id") %>%
#   mutate_at(c(3), ~replace_na(.,0))
#
# Neuroticrelateddisorders <- Neuroticrelateddisorders1 %>%
#   mutate(Neuroticrelateddisorders = Neuroticrelateddisorders.x + Neuroticrelateddisorders.y) %>%
#   mutate(Neuroticrelateddisorders = if_else(Neuroticrelateddisorders >= 1, 1, 0)) %>%
#   select(moh_study_id, Neuroticrelateddisorders)
#
# write.csv(Neuroticrelateddisorders, "Neuroticrelateddisorders.csv", row.names = FALSE)
#
# table(Neuroticrelateddisorders$Neuroticrelateddisorders)
#
# length(unique(Neuroticrelateddisorders$moh_study_id)) == nrow(Neuroticrelateddisorders)

# Neurocognitivedisorders1 <- left_join(Icd9_Neurocognitivedisorders_B, Icd10_Neurocognitivedisorders_B, by="moh_study_id") %>%
#   mutate_at(c(3), ~replace_na(.,0))
#
# Neurocognitivedisorders <- Neurocognitivedisorders1 %>%
#   mutate(Neurocognitivedisorders = Neurocognitivedisorders.x + Neurocognitivedisorders.y) %>%
#   mutate(Neurocognitivedisorders = if_else(Neurocognitivedisorders >= 1, 1, 0)) %>%
#   select(moh_study_id, Neurocognitivedisorders)
#
# write.csv(Neurocognitivedisorders, "Neurocognitivedisorders.csv", row.names = FALSE)
#
# table(Neurocognitivedisorders$Neurocognitivedisorders)
#
# length(unique(Neurocognitivedisorders$moh_study_id)) == nrow(Neurocognitivedisorders)

# Multiplementalillness1 <- left_join(Icd9_Multiplementalillness_B, Icd10_Multiplementalillness_B, by="moh_study_id") %>%
#   mutate_at(c(3), ~replace_na(.,0))
#
# Multiplementalillness <- Multiplementalillness1 %>%
# mutate(Multiplementalillness = Multiplementalillness.x + Multiplementalillness.y) %>%
# mutate(Multiplementalillness = if_else(Multiplementalillness >= 1, 1, 0)) %>%
# select(moh_study_id, Multiplementalillness)
#
# write.csv(Multiplementalillness, "Multiplementalillness.csv", row.names = FALSE)
#
# table(Multiplementalillness$Multiplementalillness)
#
# length(unique(Multiplementalillness$moh_study_id)) == nrow(Multiplementalillness)

# Intellectualdisability1 <- left_join(Icd9_Intellectualdisability_B, Icd10_Intellectualdisability_B, by="moh_study_id") %>%
# mutate_at(c(3), ~replace_na(.,.))
#
# Intellectualdisability <- Intellectualdisability1 %>%
# mutate(Intellectualdisability = Intellectualdisability.x + Intellectualdisability.y) %>%
# mutate(Intellectualdisability = if_else(Intellectualdisability >= 1, 1, 0)) %>%
# select(moh_study_id, Intellectualdisability)
#
# write.csv(Intellectualdisability, "Intellectualdisability.csv", row.names = FALSE)
#
# table(Intellectualdisability$Intellectualdisability)
#
# length(unique(Intellectualdisability$moh_study_id)) == nrow(Intellectualdisability)

# Hallucinogensuse1 <- left_join(Icd9_Hallucinogensuse_B, Icd10_Hallucinogensuse_B, by="moh_study_id") %>%
# mutate_at(c(3), ~replace_na(.,.))
#
# Hallucinogensuse <- Hallucinogensuse1 %>%
# mutate(Hallucinogensuse = Hallucinogensuse.x + Hallucinogensuse.y) %>%
# mutate(Hallucinogensuse = if_else(Hallucinogensuse >= 1, 1, 0)) %>%
# select(moh_study_id, Hallucinogensuse)
#
# write.csv(Hallucinogensuse, "Hallucinogensuse.csv", row.names = FALSE)
#
# table(Hallucinogensuse$Hallucinogensuse)
#
# length(unique(Hallucinogensuse$moh_study_id)) == nrow(Hallucinogensuse)

# Endocarditis1 <- left_join(Icd9_endocarditis_B, Icd10_Endocarditis_B, by="moh_study_id") %>%
# mutate_at(c(3), ~replace_na(.,.))
#
# Endocarditis <- Endocarditis1 %>%
# mutate(Endocarditis = Endocarditis + Endocarditis) %>%
# mutate(Endocarditis = if_else(Endocarditis >= 1, 1, 0)) %>%
# select(moh_study_id, Endocarditis)
#
# write.csv(Endocarditis, "Endocarditis.csv", row.names = FALSE)
```r
# table(Endocarditis$Endocarditis)
#
# length(unique(Endocarditis$moh_study_id)) == nrow(Endocarditis)
#
# earlyonsetdisorders1 <- left_join(Icd9_earlyonsetdisorders_B, Icd10_earlyonsetdisorders_B, by="moh_study_id") %>%
#  mutate_at(c(3), ~replace_na(.,0))
#
# earlyonsetdisorders <- earlyonsetdisorders1 %>%
#  mutate(earlyonsetdisorders = earlyonsetdisorders.x + earlyonsetdisorders.y) %>%
#  mutate(earlyonsetdisorders = if_else(earlyonsetdisorders >= 1, 1, 0)) %>%
#  select(moh_study_id, earlyonsetdisorders)
#
# write.csv(earlyonsetdisorders, "earlyonsetdisorders.csv", row.names = FALSE)
#
# table(earlyonsetdisorders$earlyonsetdisorders)
#
# length(unique(earlyonsetdisorders$moh_study_id)) == nrow(earlyonsetdisorders)
#
# Developmentdisorders1 <- left_join(Icd9_Developmentdisorders_B, Icd10_Developmentdisorders_B, by="moh_study_id") %>%
#  mutate_at(c(3), ~replace_na(.,0))
#
# Developmentdisorders <- Developmentdisorders1 %>%
#  mutate(Developmentdisorders = Developmentdisorders.x + Developmentdisorders.y) %>%
#  mutate(Developmentdisorders = if_else(Developmentdisorders >= 1, 1, 0)) %>%
#  select(moh_study_id, Developmentdisorders)
#
# write.csv(Developmentdisorders, "Developmentdisorders.csv", row.names = FALSE)
#
# table(Developmentdisorders$Developmentdisorders)
#
# length(unique(Developmentdisorders$moh_study_id)) == nrow(Developmentdisorders)
#
# Cannabinoiduse1 <- left_join(Icd9_Cannabinoiduse_B, Icd10_Cannabinoiduse_B, by="moh_study_id") %>%
#  mutate_at(c(3), ~replace_na(.,0))
#
# Cannabinoiduse <- Cannabinoiduse1 %>%
#  mutate(Cannabinoiduse = Cannabinoiduse.x + Cannabinoiduse.y) %>%
#  mutate(Cannabinoiduse = if_else(Cannabinoiduse >= 1, 1, 0)) %>%
#  select(moh_study_id, Cannabinoiduse)
#
# write.csv(Cannabinoiduse, "Cannabinoiduse.csv", row.names = FALSE)
#
# table(Cannabinoiduse$Cannabinoiduse)
#
# length(unique(Cannabinoiduse$moh_study_id)) == nrow(Cannabinoiduse)
#
# Cocaineuse1 <- left_join(Icd9_Cocaineuse_B, Icd10_Cocaineuse_B, by="moh_study_id") %>%
#  mutate_at(c(3), ~replace_na(.,0))
```
# Cocaineuse <- Cocaineuse1 %>%
#   mutate(Cocaineuse = Cocaineuse.x + Cocaineuse.y) %>%
#   mutate(Cocaineuse = if_else(Cocaineuse >= 1, 1, 0)) %>%
#   select(moh_study_id, Cocaineuse)
#
# write.csv(Cocaineuse, "Cocaineuse.csv", row.names = FALSE)
#
# table(Cocaineuse$Cocaineuse)
#
# length(unique(Cocaineuse$moh_study_id)) == nrow(Cocaineuse)

# Behaviouralpsychologicaldisturbances1 <- left_join(Icd9_Behaviouralpsychologicaldisturbances_B, Icd10_Behaviouralpsychologicaldisturbances_B, by="moh_study_id") %>%
#   mutate_at(c(3), ~replace_na(.,.0))
#
# Behaviouralpsychologicaldisturbances <- Behaviouralpsychologicaldisturbances1 %>%
#   mutate(Behaviouralpsychologicaldisturbances = Behaviouralpsychologicaldisturbances.x + Behaviouralpsychologicaldisturbances.y) %>%
#   mutate(Behaviouralpsychologicaldisturbances = if_else(Behaviouralpsychologicaldisturbances >= 1, 1, 0)) %>%
#   select(moh_study_id, Behaviouralpsychologicaldisturbances)
#
# write.csv(Behaviouralpsychologicaldisturbances, "Behaviouralpsychologicaldisturbances.csv", row.names = FALSE)
#
# table(Behaviouralpsychologicaldisturbances$Behaviouralpsychologicaldisturbances)
#
# length(unique(Behaviouralpsychologicaldisturbances$moh_study_id)) == nrow(Behaviouralpsychologicaldisturbances)

# Alcoholuse1 <- left_join(Icd9_Alcoholuse_B, Icd10_Alcoholuse_B, by="moh_study_id") %>%
#   mutate_at(c(3), ~replace_na(.,.0))
#
# Alcoholuse <- Alcoholuse1 %>%
#   mutate(Alcoholuse = Alcoholuse.x + Alcoholuse.y) %>%
#   mutate(Alcoholuse = if_else(Alcoholuse >= 1, 1, 0)) %>%
#   select(moh_study_id, Alcoholuse)
#
# write.csv(Alcoholuse, "Alcoholuse.csv", row.names = FALSE)
#
# table(Alcoholuse$Alcoholuse)
#
# length(unique(Alcoholuse$moh_study_id)) == nrow(Alcoholuse)

Data confirmation

You can also embed plots, for example:

# length(unique(substancerelateddisorders$moh_study_id))
# Mooddisorders1 <- Mooddisorders[!duplicated(Mooddisorders$moh_study_id), ]
# Mooddisorders1 <- Mooddisorders %>%
#   distinct(moh_study_id, .keep_all = TRUE)
#
# table(substancerelateddisorders$substancerelateddisorders)
# table(substancerelateddisorders1$substancerelateddisorders)
# table(substancerelateddisorders3$substancerelateddisorders)
# table(substancerelateddisorders101$substancerelateddisorders)
# table(Tobaccouse$Tobaccouse)
# table(Tobaccouse1$Tobaccouse)
# table(Mooddisorders$Mooddisorders)
# table(Mooddisorders1$Mooddisorders)
# table(ICD3$Tobaccouse)
# substancerelateddisorders2 <- unique(substancerelateddisorders [, 1:2])
# substancerelateddisorders2[duplicated(substancerelateddisorders2[,1:2])]
# table(substancerelateddisorders2$substancerelateddisorders)
# duplicated(substancerelateddisorders$substancerelateddisorders)
# write.csv(Tobaccouse1, "Tobaccouse.csv", row.names = FALSE)
# length(unique(ICD26$moh_study_id)) == nrow(ICD26)

### combining the icd datasets

# ICD1 <- left_join(substancerelateddisorders, Tobaccouse, by = "moh_study_id")
# ICD2 <- ICD1 %>%
#   distinct(moh_study_id, .keep_all = TRUE)
# ICD3 <- left_join(ICD2, Alcoholuse, by = "moh_study_id")
# ICD4 <- left_join(ICD3, Behaviouralpsychologicaldisturbances, by = "moh_study_id")
# ICD5 <- left_join(ICD4, Cannabinoiduse, by = "moh_study_id")
# ICD6 <- left_join(ICD5, Cocaineuse, by = "moh_study_id")
# ICD7 <- left_join(ICD6, Developmentdisorders, by = "moh_study_id")
# ICD8 <- left_join(ICD7, earlyonsetdisorders, by = "moh_study_id")
# ICD9 <- left_join(ICD8, Endocarditis, by = "moh_study_id")
Combining all files

```r
library(tidyverse)

## Attaching core tidyverse packages
### tidyverse 2.0.0
### v dplyr 1.1.2 v readr 2.1.4
### v forcats 1.0.0 v stringr 1.5.0
### v ggplot2 3.4.2 v tibble 3.2.1
### v lubridate 1.9.2 v tidyr 1.3.0
### v purrr 1.0.1

## Conflicts
### tidyverse_conflicts()
### x dplyr::filter() masks stats::filter()
### x dplyr::lag() masks stats::lag()
### i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```
library(dplyr)
library(gapminder)
library(caret)

## Loading required package: lattice
## Attaching package: 'caret'
## The following object is masked from 'package:purrr':
## lift

library(xgboost)

## Attaching package: 'xgboost'
## The following object is masked from 'package:dplyr':
## slice

library(forcats)
library(mice)

## Attaching package: 'mice'
## The following object is masked from 'package:stats':
## filter
## The following objects are masked from 'package:base':
## cbind, rbind

library(vctrs)

## Attaching package: 'vctrs'
## The following object is masked from 'package:dplyr':
## data_frame
## The following object is masked from 'package:tibble':
## data_frame

library(naniar)
library(arsenal)

## Attaching package: 'arsenal'
Cleaning CDR case

```r
# length(unique(vw_od_case$moh_study_id)) == nrow(vw_od_case)
#
# length(unique(vw_cdr_case$moh_study_id)) == nrow(vw_cdr_case)

# vw_cdr_case1 <- (vw_cdr_case %>%
#   mutate(across(ends_with("_date"), ~case_when( 
#       .x >= 1 ~ 1,
#   
#   TRUE ~ 0)), .names = "{.col}_b"))

# vw_cdr_case_b <- (vw_cdr_case1 %>%
#   select(moh_study_id, ends_with("_b")))

# write.csv(vw_cdr_case_b, "vw_cdr_case_b.csv", row.names = FALSE)
```

```r
# Table of data sets, descriptive

ICDderivedvariables <- read_csv("U:/Andy's Thesis/MLoverdose/MLdataset/ICDderivedvariables.csv", show_col_types = FALSE)
Pharmanetdinpin <- read_csv("U:/Andy's Thesis/MLoverdose/MLdataset/Pharmanetdinpin.csv", show_col_types = FALSE)
vw_cdr_case_b <- read_csv("U:/Andy's Thesis/MLoverdose/MLdataset/vw_cdr_case_b.csv", show_col_types = FALSE)
vw_od_case <- read_csv("U:/Andy's Thesis/MLoverdose/MLdataset/vw_od_case.csv", show_col_types = FALSE)

```
You can also embed plots, for example:

```r
# unique(length(ICDderivedvariables$moh_study_id))
# unique(length(dataset3$moh_study_id))

dataset1 <- left_join(vw_od_case, vw_cdr_case_b, by = "moh_study_id")
#
# dataset2 <- left_join(dataset1, Pharmanetdinpin_f, by = "moh_study_id")
#
# dataset3 <- left_join(dataset2, ICDderivedvariables, by = "moh_study_id")
#
# write.csv(dataset3, "datasetfirst.csv", row.names = FALSE)

dataset3 <- read_csv("U:/Andy's Thesis/MLoverdose/datasetfirst.csv", show_col_types = FALSE)
# table(Pharmanetdinpin_f$oat)
# table(Pharmanetdinpin$oat)
# Pharmanetdinpin_f <- Pharmanetdinpin %>%
#  distinct(moh_study_id, .keep_all = TRUE)
# length(unique(Pharmanetdinpin_f$moh_study_id)) == nrow(Pharmanetdinpin_f)
# table(vw_od_case$fatal_od_case)

gg_miss_var(dataset3, show_pct = TRUE)
```

---

## Warning: The `guide` argument in `scale_*()` cannot be `FALSE`. This was deprecated in ggplot2 3.3.4.
## i Please use "none" instead.
## i The deprecated feature was likely used in the naniar package.
## Please report the issue at <https://github.com/njtierney/naniar/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```r
library(tidyverse)

# Load the dataset
dataset3 <- read_csv('dataset3.csv')

# Descriptive table of the data
table(dataset3$fatal_od_case)

# Rows: 37,624
# Columns: 67

glimpse(dataset3)

# Data descriptive
```
# Pharmanetspss <- read_csv("U:/Andy's Thesis/Raw_data_used/Pharmanetspss.csv", show_col_types = FALSE)

# glimpse(Pharmanetspss)

### OST/OAT

# dinpin <- (Pharmanetspss %>%
#   mutate(across(starts_with("din_pin"), ~case_when(
#      .x=="2242963"~ 1,
#      .x=="2242964"~ 1,
#      .x=="2295695"~ 1,
#      .x=="2295709"~ 1,
#      .x=="2408104"~ 1,
#      .x=="2424851"~ 1,
#      .x=="2424878"~ 1,
#      .x=="999776"~ 1,
#      .x=="665619"~ 1,
#      .x=="655627"~ 1,
#      .x=="781460"~ 1,
#      .x=="781479"~ 1,
#      .x=="2408090"~ 1,
#      (.x >= 999792) & (.x <= 999793) ~ 1,
#      (.x >= 66999990) & (.x <= 66999993) ~ 1,
#      (.x >= 66999990) & (.x <= 66999993) ~ 1,
#      (.x >= 67000000) & (.x <= 67000004) ~ 1,
#      (.x >= 22123346) & (.x <= 22123349) ~ 1,
#      TRUE ~ 0), .names = "derivde{.col}"))

# dinpin_1 <- dinpin %>%
#   rename(ost = derivdedin_pin) %>%
#   select(moh_study_id, oat, ost)

# write.csv(dinpin_1, "Pharmanetdinpin.csv", row.names = FALSE)
D.4 ML Prediction Code for Fatal Overdose

Data preprocessing

```r
library(rlang)
library(tidyverse)

## -- Attaching core tidyverse packages --------------------------- tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ----------------------------------------------- tidyverse_conflicts() --
## x purrr::%@%()         masks rlang::%@%()
## x dplyr::filter()      masks stats::filter()
## x purrr::flatten()     masks rlang::flatten()
## x purrr::flatten_chr() masks rlang::flatten_chr()
## x purrr::flatten_dbl() masks rlang::flatten_dbl()
## x purrr::flatten_int() masks rlang::flatten_int()
## x purrr::flatten_lgl() masks rlang::flatten_lgl()
## x purrr::flatten_raw() masks rlang::flatten_raw()
## x purrr::invoke()      masks rlang::invoke()
## x dplyr::lag()         masks stats::lag()
## x purrr::splice()      masks rlang::splice()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(dplyr)
library(gapminder)
library(xgboost)

##
## Attaching package: 'xgboost'
##
## The following object is masked from 'package:dplyr':
##
## slice

library(forcats)
library(mice)

##
## Attaching package: 'mice'
##
## The following object is masked from 'package:stats':
##
## filter
##
## The following objects are masked from 'package:base':
```
## cbind, rbind

```r
library(vctrs)
```

## Attaching package: 'vctrs'

## The following object is masked from 'package:plyr':

## data_frame

## The following object is masked from 'package:tibble':

## data_frame

```r
library(naniar)
library(yardstick)
```

## For binary classification, the first factor level is assumed to be the event.
## Use the argument `event_level = "second"` to alter this as needed.

## Attaching package: 'yardstick'

## The following object is masked from 'package:readr':

## spec

```r
library(forcats)
library(Boruta)
library(ROSE)
```

## Loaded ROSE 0.0-4

```r
library(imbalance)
library(ROCR)
library(mlbench)
library(ggplot2)
library(data.table)
```

## Attaching package: 'data.table'

## The following objects are masked from 'package:lubridate':

## hour, isoweek, mday, minute, month, quarter, second, wday, week,

## yday, year

## The following objects are masked from 'package:plyr':

## between, first, last

##
library(here)
## here() starts at R:/atai-22-075/Andy's Thesis/MLoverdose
library(caret)
## Loading required package: lattice
## Attaching package: 'caret'
## The following objects are masked from 'package:yardstick':
##   precision, recall, sensitivity, specificity
## The following object is masked from 'package:purrr':
##   lift
dataset3 <- read_csv("U:/Andy's Thesis/MLoverdose/datasetfirst.csv" , show_col_types = FALSE)

# Control <- dataset3 %>%
#   filter(fatal_od_case == 0)
# # Case <- dataset3 %>%
#   filter(fatal_od_case == 1)
# colSums(is.na(dataset3))
# # colSums(is.na(Case))
# # colSums(is.na(Control))
gg_miss_var(dataset3, show_pct = TRUE)

## Warning: The `guide` argument in `scale_*()` cannot be `FALSE`. This was deprecated in
## ggplot2 3.3.4.
## i Please use "none" instead.
## i The deprecated feature was likely used in the naniar package.
## i Please report the issue at <https://github.com/njtierney/naniar/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
Removing osteo_fracture as there are none in this population “osteo_fracture_date_b”

```r
data1 <- dataset3 %>%
  select(moh_study_id, ost, rheumatoid_arthritis_date_b, parkinsonism_date_b, osteo_arthritis_date_b, osteo_porosis_date_b, ms_date_b, mood_anx_date_b, isch_stroke_date_b, ihd_date_b, hypertension_date_b, hosp_tia_date_b, hosp_stroke_date_b, heart_failure_date_b, haemorrh_stroke_date_b, osteo_fracture_date_b, diabetes_date_b, depression_api_date_b, depression_date_b, copd_date_b, ckd_date_b, asthma_api_date_b, alpina_date_b, smf_api_date_b, smf_date_b)

data2 <- data1 %>%
  distinct(moh_study_id, .keep_all = TRUE)
```

314
data3 <- data2[-1]

moh <- data1 %>% select(moh_study_id)

table(data3$fatal_od_case)

##     0     1
## 30253  6426

##splitting

set.seed(2001)

na_count0 <- sapply(data3, function(y) sum(length(which(is.na(y)))))

na_count00 <- data.frame(na_count0)

na_count00

##             na_count0
## ost           3306
## rheumatoid_arthritis_date_b 5175
## parkinsonism_date_b           5175
## osteo_arthritis_date_b       5175
## osteo_porosis_date_b         5175
## ms_date_b                   5175
## mood_anx_date_b              5175
## isch_stroke_date_b           5175
## ihd_date_b                  5175
## hypertension_date_b          5175
## hosp_tia_date_b              5175
## hosp_stroke_date_b           5175
## heart_failure_date_b         5175
## haemorr_stroke_date_b        5175
## epilepsy_date_b              5175
## diabetes_date_b              5175
## depression_date_b            5175
## copd_date_b                  5175
## ckd_date_b                   5175
## asthma_date_b                5175
## angina_date_b                5175
## ami_date_b                   5175
## alzheimer_dementia_date_b    5175
## Tobaccouse                   1735
## Tissueinfection             1735
## substancerelateddisorders.x  1735
## Stimulantuse                 1735
## Sepsis                       1735
## Sedativeandhypnoticuse       1735
## psychoticdisorders           1735
## Polysubstance 1735
## Personalitydisorders 1735
## Otherpsychoactivedruguse 1735
## Osteomyelitis 1735
## Opioiduse 1735
## Neuroticrelateddisorders 1735
## Neurocognitivedisorders 1735
## Multiplementalillness 1735
## Mooddisorders 1735
## Intellectualdisability 1735
## Hallucinogensuse 1735
## Endocarditis 1735
## earlyonsetdisorders 1735
## Developmentdisorders 1735
## Cocaineuse 1735
## Cannabinoiduse 1735
## Behaviouralpsychologicaldisturbances 1735
## Alcoholuse 1735
## fatal_od_case 0
## total_od_n 0
## first_od_date 0
## last_od_date 0

data <- data3

index = createDataPartition(y = data$fatal_od_case, p = 0.7, list = FALSE)

train.set = data[index,]
test.set = data[-index,]

gg_miss_var(train.set)
```r
na_count1 <- sapply(train.set, function(y) sum(length(which(is.na(y)))))
```
na_count <- data.frame(na_count1)

na_count

###       na_count1
### ost      2306
### rheumatoid_arthritis_date_b       3611
### parkinsonism_date_b                3611
### osteo_arthritis_date_b             3611
### osteo_porosis_date_b               3611
### ms_date_b                           3611
### mood_anx_date_b                     3611
### isch_stroke_date_b                  3611
### ihd_date_b                          3611
### hypertension_date_b                 3611
### hosp_tia_date_b                     3611
### hosp_stroke_date_b                  3611
### heart_failure_date_b                3611
### haemorr_stroke_date_b               3611
### epilepsy_date_b                     3611
### diabetes_date_b                     3611
### depression_date_b                   3611
### copd_date_b                         3611
### ckd_date_b                          3611
### asthma_date_b                       3611
### angina_date_b                       3611
### ami_date_b                          3611
### alzheimer_dementia_date_b            3611
### Tobaccouse                           1190
### Tissueinfection                      1190
### substancerelateddisorders.x          1190
### Stimulantuse                         1190
### Sepsis                                1190
### Sedativeandhypnoticuse               1190
### psychoticdisorders                   1190
### Polysubstance                        1190
### Personalitydisorders                 1190
### Otherpsychoactivedruguse             1190
### Osteomyelitis                         1190
### Opioiduse                             1190
### Neuroticrelateddisorders             1190
### Neurocognitivedisorders              1190
### Multiplementalillness                1190
### Mooddisorders                        1190
### Intellectualdisability               1190
### Hallucinogensuse                     1190
### Endocarditis                          1190
### earlyonsetdisorders                  1190
### Developmentdisorders                 1190
### Cocaineuse                           1190
### Cannabinoiduse                       1190
## Behavioural psychological disturbances

### Alcohol use

```r
na_count2 <- sapply(test.set, function(y) sum(length(which(is.na(y)))))
```

```r
na_count3 <- data.frame(na_count2)
```

```r
na_count3
```

```r
## na_count2
## ost 1000
## rheumatoid_arthritis_date_b 1564
## parkinsonism_date_b 1564
## osteo_arthritis_date_b 1564
## osteo_porosis_date_b 1564
## ms_date_b 1564
## mood_anx_date_b 1564
## isch_stroke_date_b 1564
## ihd_date_b 1564
## hypertension_date_b 1564
## hosp_tia_date_b 1564
## hosp_stroke_date_b 1564
## heart_failure_date_b 1564
## haemorr_stroke_date_b 1564
## epilepsy_date_b 1564
## diabetes_date_b 1564
## depression_date_b 1564
## copd_date_b 1564
## ckd_date_b 1564
## asthma_date_b 1564
## angina_date_b 1564
## ami_date_b 1564
## alzheimer_dementia_date_b 1564
## Tobaccouse 545
## Tissueinfection 545
## substancerelateddisorders.x 545
## Stimulantuse 545
## Sepsis 545
## Sedativeandhypnoticuse 545
## psychoticdisorders 545
## Polysubstance 545
## Personalitydisorders 545
## Otherpsychoactivedruguse 545
## Osteomyelitis 545
## Opioiduse 545
## Neuroticrelateddisorders 545
## Neurocognitivedisorders 545
```
## Multiplementalillness                      545
## Mooddisorders                              545
## Intellectualdisability                     545
## Hallucinogensuse                           545
## Endocarditis                               545
## earlyonsetdisorders                        545
## Developmentdisorders                       545
## Cocaineuse                                 545
## Cannabinoiduse                             545
## Behaviouralpsychologicaldisturbances       545
## Alcoholuse                                 545
## fatal_od_case                                0
## total_od_n                                   0
## first_od_date                                0
## last_od_date                                 0

fatal <- data %>%
  filter(fatal_od_case == 1 )

control <- data %>%
  filter(fatal_od_case == 0 )

na_count1 <- sapply(fatal, function(y) sum(length(which(is.na(y)))))

na_count <- data.frame(na_count1)

na_count

## na_count1
##        ost   684
## rheumatoid_arthritis_date_b             939
## parkinsonism_date_b                     939
## osteo_arthritis_date_b                  939
## osteo_porosis_date_b                    939
## ms_date_b                               939
## mood_anx_date_b                         939
## isch_stroke_date_b                      939
## ihd_date_b                              939
## hypertension_date_b                     939
## hosp_tia_date_b                         939
## hosp_stroke_date_b                      939
## heart_failure_date_b                    939
## haemorrh_stroke_date_b                  939
## epilepsy_date_b                         939
## diabetes_date_b                         939
## depression_date_b                       939
## copd_date_b                             939
## ckd_date_b                              939
## asthma_date_b                           939
## angina_date_b                           939
na_count2 <- sapply(control, function(y) sum(length(which(is.na(y)))))

na_count3 <- data.frame(na_count2)

na_count3

```r
##                 na_count2
## ost               2622
## rheumatoid_arthritis_date_b  4236
## parkinsonism_date_b         4236
## osteo_arthritis_date_b      4236
## osteo_porosis_date_b        4236
## ms_date_b                4236
## mood_anx_date_b           4236
## isch_stroke_date_b         4236
## ihd_date_b                4236
## hypertension_date_b        4236
## hosp_tia_date_b           4236
## hosp_stroke_date_b         4236
```
## heart_failure_date_b 4236
## haemorr_stroke_date_b 4236
## epilepsy_date_b 4236
## diabetes_date_b 4236
## depression_date_b 4236
## copd_date_b 4236
## ckd_date_b 4236
## asthma_date_b 4236
## angina_date_b 4236
## ami_date_b 4236
## alzheimer_dementia_date_b 4236
## tobaccouse 1398
## tissueinfection 1398
## substancerelateddisorders.x 1398
## stimulantuse 1398
## sepsis 1398
## sedativeandhypnoticuse 1398
## psychoticdisorders 1398
## polysubstance 1398
## personalitydisorders 1398
## otherpsychoactivedruguse 1398
## osteomyelitis 1398
## opioiduse 1398
## neuroticrelateddisorders 1398
## neurocognitivedisorders 1398
## multiplementalillness 1398
## mooddisorders 1398
## intellectualdisability 1398
## hallucinogensuse 1398
## endocarditis 1398
## earlyonsetdisorders 1398
## developmentdisorders 1398
## cocaineuse 1398
## cannabinoiduse 1398
## behaviouralpsychologicaldisturbances 1398
## alcoholuse 1398
## fatal_od_case 0
## total_od_n 0
## first_od_date 0
## last_od_date 0

# test.set$fatal_od_case = as.factor(test.set$fatal_od_case)
# train.set$fatal_od_case = as.factor(train.set$fatal_od_case)
#
# dim(train.set)
# dim(test.set)
#
# table(train.set$fatal_od_case)
# table(test.set$fatal_od_case)
```r
# table(data3$fatal_od_case)
##
##     0     1
## 30253 6426

# table(test.set$fatal_od_case)
##
##     0     1
## 9041 1962

# table(train.set$fatal_od_case)
##
##     0     1
## 21212 4464

# write.csv(test.set1, "test.set.csv", row.names = FALSE)
# write.csv(train.set1, "train.set.csv", row.names = FALSE)

train.set1 <- read_csv("U:/Andy's Thesis/MLoverdose/train.set.csv", show_col_types = FALSE)
test.set1 <- read_csv("U:/Andy's Thesis/MLoverdose/test.set.csv", show_col_types = FALSE)

gg_miss_var(train.set1)
```
gg_miss_var(test.set1)
table(train.set1$fatal_od_case)

##
## 0 1
## 21212 4464
##

table(test.set1$fatal_od_case)

##
## 0 1
## 9041 1962
##

Class balance

library(keras)
library(tensorflow)
library(rlang)
library(tidyverse)

### -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
### v dplyr 1.1.2 v readr 2.1.4
### v forcats 1.0.0 v stringr 1.5.0
### v ggplot2 3.4.2 v tibble 3.2.1
### v lubridate 1.9.2 v tidyr 1.3.0
### v purrr 1.0.1
### -- Conflicts ------------------------------------------
tidyverse_conflicts()
### x purrr::%@%() masks rlang::%@%()
### x dplyr::filter() masks stats::filter()
### x purrr::flatten() masks rlang::flatten()
### x purrr::flatten_chr() masks rlang::flatten_chr()
### x purrr::flatten_dbl() masks rlang::flatten_dbl()
### x purrr::flatten_int() masks rlang::flatten_int()
### x purrr::flatten_lgl() masks rlang::flatten_lgl()
### x purrr::flatten_raw() masks rlang::flatten_raw()
### x purrr::invoke() masks rlang::invoke()
### x dplyr::lag() masks stats::lag()
### x purrr::splice() masks rlang::splice()
### i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(dplyr)
library(gapminder)
library(xgboost)

###
### Attaching package: 'xgboost'
###
### The following object is masked from 'package:dplyr':
###
###   slice

library(forcats)
library(mice)

###
### Attaching package: 'mice'
###
### The following object is masked from 'package:stats':
###
###   filter
###
### The following objects are masked from 'package:base':
###
###   cbind, rbind

library(vctrs)

###
### Attaching package: 'vctrs'
The following object is masked from 'package:delayed':

data_frame

The following object is masked from 'package:tidyverse':

data_frame

library(naniar)
library(yardstick)

For binary classification, the first factor level is assumed to be the event.
Use the argument `event_level = "second"` to alter this as needed.

Attaching package: 'yardstick'

The following object is masked from 'package:readr':

spec

The following object is masked from 'package:keras':

get_weights

library(forcats)
library(Boruta)
library(ROSE)

Loaded ROSE 0.0-4

library(imbalance)
library(rn)
library(ROCR)
library(SHAPforxgboost)
library(randomForest)

randomForest 4.6-14
Type rfNews() to see new features/changes/bug fixes.

Attaching package: 'randomForest'

The following object is masked from 'package:delayed':

combine

The following object is masked from 'package:ggplot2':

margin
library(mlbench)
library(ggplot2)
library(data.table)

##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:lubridate':
##
## hour, isoweek, mday, minute, month, quarter, second, wday, week,
## yday, year
##
## The following objects are masked from 'package:dplyr':
##
## between, first, last
##
## The following object is masked from 'package:purrr':
##
## transpose
##
## The following object is masked from 'package:rlang':
##
## :=

library(here)

## here() starts at R:/atai-22-075/Andy's Thesis/MLoverdose

library(tidymodels)

##-- Attaching packages ----------------- tidymodels 0.1.1 --
## v broom 1.0.4 v recipes 0.1.14
## v dials 0.0.9 v rsample 0.0.8
## v infer 0.5.3 v tune 0.1.1
## v modeldata 0.1.0 v workflows 0.2.1
## v parsnip 0.1.4
##-- Conflicts ------------------------- tidymodels_conflicts() --
## x purrr::%@%() masks rlang::%@%()
## x data.table:::=() masks rlang:::=()
## x data.table::between() masks dplyr::between()
## x randomForest::combine() masks dplyr::combine()
## x vctrs::data_frame() masks dplyr::data_frame(), tibble::data_frame()
## x scales::discard() masks purrr::discard()
## x mice::filter() masks dplyr::filter(), stats::filter()
## x data.table::first() masks dplyr::first()
## x recipes::fixed() masks stringr::fixed()
## x purrr::flatten() masks rlang::flatten()
## x purrr::flatten_chr() masks rlang::flatten_chr()
## x purrr::flatten_dbl() masks rlang::flatten_dbl()
## x purrr::flatten_int() masks rlang::flatten_int()
## x purrr::flatten_lgl() masks rlang::flatten_lgl()
## x purrr::flatten_raw() masks rlang::flatten_raw()
library(parsnip)
library(caret)

## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following objects are masked from 'package:yardstick':
##
##     precision, recall, sensitivity, specificity
##
## The following object is masked from 'package:purrr':
##
##     lift
##
## The following object is masked from 'package:tensorflow':
##
##     train

library(kknn)

##
## Attaching package: 'kknn'
##
## The following object is masked from 'package:caret':
##
##     contr.dummy

library(yardstick)
library(tune)
library(MatchIt)

train.set <- read_csv("U:/Andy's Thesis/MLoverdose/train.set.csv" , show_col_types = FALSE)
test.set <- read_csv("U:/Andy's Thesis/MLoverdose/test.set.csv" , show_col_types = FALSE)
table(train.set$fatal_od_case)

##
##   0 1
## 21212 4464
### Split Data

### 3:1
undersample_rose_train_1 <- read_csv("U:/Andy's Thesis/MLoverdose/undersample_rose_train_1.csv", show_col_types = FALSE)
	able(undersample_rose_train_1$fatal_od_case)

##
##     0     1
## 12039  4464

# undersample_rose_train_2 <- ovun.sample(formula = fatal_od_case ~ ., data = train.set, method = "under", p = 0.27, seed = 2001)
#
#
# undersample_rose_train_1 <- undersample_rose_train_2$data
#
#
# table(undersample_rose_train_1$fatal_od_case)
#
write.csv(undersample_rose_train_1, "undersample_rose_train_1.csv", row.names = FALSE)

### 1:1
undersample_rose_train_2 <- read_csv("U:/Andy's Thesis/MLoverdose/undersample_rose_train_2.csv", show_col_types = FALSE)
	able(undersample_rose_train_2$fatal_od_case)

##
##    0    1
## 4453 4464

# undersample_rose_train_2 <- ovun.sample(formula = fatal_od_case ~ ., data = train.set, method = "under", p = 0.5, seed = 2001)
#
#
# undersample_rose_train_2 <- undersample_rose_train_2$data
#
#
# table(undersample_rose_train_2$fatal_od_case)
#
write.csv(undersample_rose_train_2, "undersample_rose_train_2.csv", row.names = FALSE)

### bothsample

bothsample_rose_train_1 <- read_csv("U:/Andy's Thesis/MLoverdose/bothsample_rose_train_1.csv", show_col_types = FALSE)
	able(bothsample_rose_train_1$fatal_od_case)
ML modeling with Caret

```r
library(keras)
library(tensorflow)
library(rlang)
library(tidyverse)
library(dplyr)
library(gapminder)
library(caret)
library(xgboost)
library(forcats)
library(mice)
library(vctrs)
```
library(naniar)
library(yardstick)
library(forcats)
library(Boruta)
library(ROSE)
library(imbalance)
library(rnn)
library(ROCR)
library(SHAPforxgboost)
library(randomForest)
library(mlbench)
library(ggplot2)
library(data.table)
library(here)
library(pROC)
library(ROCR)

### Testing the original sample

datasettrain <- read_csv("U:/Andy's Thesis/MLoverdose/train.set.csv", show_col_types = FALSE)
datasettest <- read_csv("U:/Andy's Thesis/MLoverdose/test.set.csv", show_col_types = FALSE)
datasettrain <- datasettrain %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))
datasettest <- datasettest %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))

train.set <- datasettrain
test.setog <- datasettest

dim(train.set)
## [1] 25676   50
dim(test.setog)
## [1] 11003   50

nrow(train.set)
## [1] 25676

train.set = train.set %>% mutate_if(is.character, as.factor)
test.setog = test.setog %>% mutate_if(is.character, as.factor)

class(train.set$fatal_od_case)
## [1] "factor"
class(test.setog$fatal_od_case)
## [1] "factor"

table(train.set$fatal_od_case)

##      No   Yes
## 21212  4464

grid_default <- expand.grid(
  nrounds = 300,
  max_depth = 5,
  eta = 0.3, gamma = 0.001,
  colsample_bytree = 0.75,
  min_child_weight = 1,
  subsample = 1
)

train_control <- caret::trainControl(
  method = "none",
  search = "grid",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
  allowParallel = TRUE
)

model = caret::train(fatal_od_case ~ .,
  data = train.set,
  trControl = train_control,
  method = "xgbTree",
  tuneGrid = grid_default,
  verbose = TRUE
)

pred <- predict(model, test.set$og)

confusionMatrix(pred, test.set$og$fatal_od_case, positive = "Yes")
## Kappa : 0.0379
## Mcnemar's Test P-Value : <2e-16
## Sensitivity : 0.044852
## Specificity : 0.980201
## Pos Pred Value : 0.329588
## Neg Pred Value : 0.825447
## Prevalence : 0.178315
## Detection Rate : 0.007998
## Detection Prevalence : 0.024266
## Balanced Accuracy : 0.512527
## 'Positive' Class : Yes
##
## `saveRDS(model, file=".rds")`
##
```r
class(train.set$fatal_od_case)
## [1] "factor"
```
## Random Forest

grid_default <- expand.grid(
  .mtry = c(1:7),
  .splitrule="gini",
  .min.node.size=c(1)
)

rf2 = caret::train(fatal_od_case ~ .,
  data = train.set,
  method = "ranger",
  tuneGrid = grid_default,
  verbose = TRUE
)

P <- predict(rf2, test.setog)
confusionMatrix(P, test.setog$fatal_od_case, positive = "Yes")
## Confusion Matrix and Statistics
## Reference
## Prediction No Yes
## No 9040 1952
## Yes 1 10
## Accuracy : 0.8225
## 95% CI : (0.8152, 0.8296)
## No Information Rate : 0.8217
## P-Value [Acc > NIR] : 0.4172
##
## Kappa : 0.0082
##
## McNemar's Test P-Value : <2e-16
##
## Sensitivity : 0.0050968
## Specificity : 0.9998894
## Pos Pred Value : 0.9090909
## Neg Pred Value : 0.8224163
## Prevalence : 0.1783150
## Detection Rate : 0.0009088
## Detection Prevalence : 0.0009997
## Balanced Accuracy : 0.5024931
##
## 'Positive' Class : Yes
##
### Testing Different Samples
### Undersample 3:1
undersample_rose_train_1 <- read_csv("U:/Andy's Thesis/MLoverdose/undersample_rose_train_1.csv", show_col_types = FALSE)

undersample_rose_train_1 <- undersample_rose_train_1 %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))

train.set <- undersample_rose_train_1

train.set = train.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 16503 50

nrow(train.set)
## [1] 16503

table(train.set$fatal_od_case)
##
## No Yes
## 12039 4464

grid_default <- expand.grid(
  nrounds=300,
  max_depth=5,
  eta=0.3, gamma=0.001,
  colsample_bytree=0.75,
  min_child_weight=1,
  subsample=1)

train_control <- caret::trainControl(
method = "none",
search = "grid",
summaryFunction = twoClassSummary,
returnResamp = "none",
classProbs = TRUE,
savePredictions = TRUE,
verboseIter = FALSE,
allowParallel = TRUE
)

model = caret::train(fatal_od_case ~ .,
data = train.set,
trControl = train_control,
method = "xgbTree",
tuneGrid = grid_default,
verbose = TRUE
)
pred <- predict(model, test.setog)
classProbs = TRUE,
verboseIter = FALSE,
allowParallel = TRUE
)

model = caret::train(fatal_od_case ~ .,
data = train.set,
trControl = train_control,
method = "xgbTree",
tuneGrid = grid_default,
verbose = TRUE
)
pred <- predict(model, test.setog)
confusionMatrix(pred, test.setog$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference  
## Prediction  No  Yes
## No  8364 1666
## Yes  677  296
##
## Accuracy : 0.7871
## 95% CI : (0.7793, 0.7947)
## No Information Rate : 0.8217
## P-Value [Acc > NIR] : 1
##
## Kappa : 0.0947
##
## Mcnemar's Test P-Value : <2e-16
##
## Sensitivity : 0.15087
## Specificity : 0.92512
## Pos Pred Value : 0.30421
## Neg Pred Value : 0.83390
## Prevalence : 0.17832
## Detection Rate : 0.02690
## Detection Prevalence : 0.08843
## Balanced Accuracy : 0.53799
##
## 'Positive' Class : Yes
##
##saveRDS(model, file = ".rds")

#Random Forest
grid_default <- expand.grid(
  .mtry = c(1:7),
  .splitrule="gini",
  .min.node.size=c(1))
)

rf1 = caret::train(fatal_od_case ~ .,
  data = train.set,
  method = "ranger",
  tuneGrid = grid_default,
  verbose = TRUE
)

pred <- predict(rf1, test.setog)

confusionMatrix(pred, test.setog$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference          
## Prediction  No  Yes
##       No  9019 1932
##       Yes   22   30
##
##                Accuracy : 0.8224
##                  95% CI : (0.8151, 0.8295)
##     No Information Rate : 0.8217
##     P-Value [Acc > NIR] : 0.4269
##
##                   Kappa : 0.0208
##
##   Mcnemar's Test P-Value : <2e-16
##
##   Sensitivity : 0.015291
##   Specificity : 0.997567
##  Pos Pred Value : 0.576923
##  Neg Pred Value : 0.823578
##    Prevalence : 0.178315
## Detection Rate : 0.002727
## Detection Prevalence : 0.004726
## Balanced Accuracy : 0.506429
##
## 'Positive' Class : Yes
##
##
####Undersample 1:1

undersample_rose_train_2 <- read_csv("U:/Andy's Thesis/MLoverdose/undersample_rose_train_2.csv", show_col_types = FALSE)

undersample_rose_train_2 <- undersample_rose_train_2 %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))
train.set <- undersample_rose_train_2

train.set = train.set %>% mutate_if(is.character, as.factor)

\textbf{dim}(train.set)

\#
\#
[1] 8917 50

\textbf{nrow}(train.set)

\#
\#
[1] 8917

\textbf{table}(train.set$fatal_od_case)

\#
\#
No Yes
4453 4464

\textbf{grid_default} <- \textbf{expand.grid}(nrounds= 300,
max_depth = 5,
eta = 0.3, gamma = 0.001,
colsample_bytree = 0.75,
min_child_weight = 1,
subsample = 1)

\textbf{train_control} <- caret::\textbf{trainControl}(method = "none",
search = "grid",
summaryFunction = twoClassSummary,
returnResamp = "none",
classProbs = TRUE,
savePredictions = TRUE,
verboseIter = FALSE,
allowParallel = TRUE)

\textbf{model} = caret::\textbf{train}(fatal_od_case ~ .,
data = train.set,
trControl = train_control,
method = "xgbTree",
tuneGrid = grid_default,
verbose = TRUE)

\textbf{pred} <- \textbf{predict}(model, test.setog)

\textbf{confusionMatrix}(pred, test.setog$fatal_od_case, positive = "Yes")

\#
\#
Confusion Matrix and Statistics
\#
\#
Reference
## Prediction  No  Yes
##        No  5275  779
##        Yes 3766 1183
##
##                Accuracy : 0.5869
##                  95% CI : (0.5777, 0.5962)
##     No Information Rate : 0.8217
##     P-Value [Acc > NIR] : 1
##
##                Kappa : 0.1168
##
## McNemar's Test P-Value : <2e-16
##
##          Sensitivity : 0.6030
##          Specificity : 0.5835
##        Pos Pred Value : 0.2390
##        Neg Pred Value : 0.8713
##          Prevalence : 0.1783
##      Detection Rate : 0.1075
## Detection Prevalence : 0.4498
## Balanced Accuracy : 0.5932
##
## 'Positive' Class : Yes
##
## #saveRDS(model, file=".rds")
##
## #Random Forest

grid_default <- expand.grid(.mtry = c(1:7), .splitrule="gini", .min.node.size=c(1))
rf1 = caret::train(fatal_od_case ~ ., data = train.set, method = "ranger", tuneGrid = grid_default, verbose = TRUE)
pred <- predict(rf1, test.setog)
confusionMatrix(pred, test.setog$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference
## Prediction  No  Yes
##        No  5550  786
##        Yes 3491 1176
##
## Accuracy : 0.6113
## 95% CI : (0.6021, 0.6204)
## No Information Rate : 0.8217
## P-Value [Acc > NIR] : 1
## Kappa : 0.1385
## McNemar's Test P-Value : <2e-16
## Sensitivity : 0.5994
## Specificity : 0.6139
## Pos Pred Value : 0.2520
## Neg Pred Value : 0.8759
## Prevalence : 0.1783
## Detection Rate : 0.1069
## Detection Prevalence : 0.4242
## Balanced Accuracy : 0.6066
## 'Positive' Class : Yes

## Both sample 1:1

bothsample_rose_train_1 <- read_csv("U:/Andy's Thesis/MLoverdose/bothsample_rose_train_1.csv", show_col_types = FALSE)

bothsample_rose_train_1 <- bothsample_rose_train_1 %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))

train.set <- bothsample_rose_train_1

train.set = train.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 25676 50

nrow(train.set)
## [1] 25676

table(train.set$fatal_od_case)
##
## No Yes
## 12876 12800

grid_default <- expand.grid(
nrounds = 300,
  max_depth = 5,
  eta = 0.3, gamma = 0.001,
  colsample_bytree = 0.75,
  min_child_weight = 1,
subsample = 1
)

train_control <- caret::trainControl(
    method = "none",
    search = "grid",
    summaryFunction = twoClassSummary,
    returnResamp = "none",
    classProbs = TRUE,
    savePredictions = TRUE,
    verboseIter = FALSE,
    allowParallel = TRUE
)

model = caret::train(fatal_od_case ~ .,
    data = train.set,
    trControl = train_control,
    method = "xgbTree",
    tuneGrid = grid_default,
    verbose = TRUE
)

pred <- predict(model, test.setog)
confusionMatrix(pred, test.setog$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
## Reference
## Prediction  No  Yes
##        No  5973  907
##        Yes 3068 1055
##
## Accuracy : 0.6387
## 95% CI : (0.6297, 0.6477)
## No Information Rate : 0.8217
## P-Value [Acc > NIR] : 1
##
## Kappa : 0.1386
##
## Mcnemar's Test P-Value : <2e-16
##
## Sensitivity : 0.53772
## Specificity : 0.66066
## Pos Pred Value : 0.25588
## Neg Pred Value : 0.86817
## Prevalence : 0.17832
## Detection Rate : 0.09588
## Detection Prevalence : 0.37472
## Balanced Accuracy : 0.59919
##
##
## 'Positive' Class : Yes
##
#saveRDS(model, file="rds")

#Random Forest

grid_default <- expand.grid(
  .mtry = c(1:7),
  .splitrule="gini",
  .min.node.size=c(1)
)

rf1 = caret::train(fatal_od_case ~ .,
  data = train.set,
  method = "ranger",
  tuneGrid = grid_default,
  verbose = TRUE
)

pred <- predict(rf1, test.setog)
confusionMatrix(pred, test.setog$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference
## Prediction  No Yes
##    No   6366 1020
##    Yes  2675  942
##
##                Accuracy : 0.6642
##                  95% CI : (0.6553, 0.673)
##     No Information Rate : 0.8217
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1385
##
##  Mcnemar's Test P-Value : <2e-16
##
## Sensitivity : 0.48012
## Specificity : 0.70413
## Pos Pred Value : 0.26044
## Neg Pred Value : 0.86190
## Prevalence : 0.17832
## Detection Rate : 0.08561
## Detection Prevalence : 0.32873
## Balanced Accuracy : 0.59212
##
## 'Positive' Class : Yes
##
##Oversample 1:1
```r
oversample_rose_train_1 <- read_csv("U:/Andy's Thesis/MLoverdose/oversample_rose_train_1.csv", show_col_types = FALSE)

oversample_rose_train_1 <- oversample_rose_train_1 %>%
mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))

train.set <- oversample_rose_train_1

train.set = train.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 42230   50

nrow(train.set)
## [1] 42230

table(train.set$fatal_od_case)
##
##    No   Yes
## 21212 21018

grid_default <- expand.grid(
  nrounds = 300,
  max_depth = 5,
  eta = 0.3, gamma = 0.001,
  colsample_bytree = 0.75,
  min_child_weight = 1,
  subsample = 1)

train_control <- caret::trainControl(
  method = "none",
  search = "grid",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
  allowParallel = TRUE)

model <- caret::train(fatal_od_case ~ ., 
  data = train.set,
  trControl = train_control,
  method = "xgbTree",
  tuneGrid = grid_default,
  verbose = TRUE)
```
pred <- predict(model, test.setog)
confusionMatrix(pred, test.setog$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference
## Prediction  No  Yes
##        No  6253  990
##        Yes 2788  972
##
##             Accuracy : 0.6566
## 95% CI : (0.6477, 0.6655)
## No Information Rate : 0.8217
## P-Value [Acc > NIR] : 1
##
##    Kappa : 0.1377
##
## Mcnemar's Test P-Value : <2e-16
##
## Sensitivity : 0.49541
## Specificity : 0.69163
## Pos Pred Value : 0.25851
## Neg Pred Value : 0.86332
## Prevalence : 0.17832
## Detection Rate : 0.08834
## Detection Prevalence : 0.34172
## Balanced Accuracy : 0.59352
##
## 'Positive' Class : Yes
##
#saveRDS(model, file=".rds")

#Random Forest

ggrid_default <- expand.grid(  
  .mtry = c(1:7),
  .splitrule="gini",
  .min.node.size=c(1)
)

rf1 = caret::train(fatal_od_case ~ .,  
  data = train.set,  
  method = "ranger",  
  tuneGrid = grid_default,  
  verbose = TRUE
)

pred <- predict(rf1, test.setog)
confusionMatrix(pred, test.setog$fatal_od_case, positive = "Yes")
## Confusion Matrix and Statistics

### Reference
### Prediction   No  Yes
##   No  6682 1109
##   Yes 2359  853

### Accuracy : 0.6848
### 95% CI : (0.676, 0.6935)
### No Information Rate : 0.8217
### P-Value [Acc > NIR] : 1
### Kappa : 0.1391

### McNemar's Test P-Value : <2e-16
### Sensitivity : 0.43476
### Specificity : 0.73908
### Pos Pred Value : 0.26557
### Neg Pred Value : 0.85766
### Prevalence : 0.17832
### Detection Rate : 0.07752
### Detection Prevalence : 0.29192
### Balanced Accuracy : 0.58692

'Positive' Class : Yes

###feature selected (From Boruta)

```r
bothsample_rose_train_1 <- read_csv("U:/Andy's Thesis/MLoverdose/bothsample_rose_train_1.csv", show_col_types = FALSE)
bothsample_rose_train_1 <- bothsample_rose_train_1 %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))
train.set <- bothsample_rose_train_1
train.set = train.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 25676  50
nrow(train.set)
## [1] 25676
table(train.set$fatal_od_case)
##
##   No  Yes
## 12876 12800
```
train.set <- train.set %>%
  select(fatal_od_case, total_od_n, alzheimer_dementia_date_b, ami_date_b, angina_date_b, ckd_date_b, copd_date_b, depression_date_b, diabetes_date_b, heart_failure_date_b, hypertension_date_b, hosp_stroke_date_b, haemorrh_stroke_date_b, isch_stroke_date_b, ihd_date_b, mood_anx_date_b, osteo_arthritis_date_b, ost, substancerelateddisorders.x, Tobaccouse, Cocaineuse, earlyonsetdisorders, Mooddisorders, Multiplealillness, Neurocognitivedisorders, Neuroticrelateddisorders, Opioiduse, Osteomyelitis, Otherpsychosocialdruguse, Personalitydisorders, Polysubstance, psychoticdisorders, Sepsis, Stimulantuse, Tissueinfection)

test.set <- test.setog %>%
  select(fatal_od_case, total_od_n, alzheimer_dementia_date_b, ami_date_b, angina_date_b, ckd_date_b, copd_date_b, depression_date_b, diabetes_date_b, heart_failure_date_b, hypertension_date_b, hosp_stroke_date_b, haemorrh_stroke_date_b, isch_stroke_date_b, ihd_date_b, mood_anx_date_b, osteo_arthritis_date_b, ost, substancerelateddisorders.x, Tobaccouse, Cocaineuse, earlyonsetdisorders, Mooddisorders, Multiplealillness, Neurocognitivedisorders, Neuroticrelateddisorders, Opioiduse, Osteomyelitis, Otherpsychosocialdruguse, Personalitydisorders, Polysubstance, psychoticdisorders, Sepsis, Stimulantuse, Tissueinfection)

grid_default <- expand.grid(
  nrounds = 300,
  max_depth = 5,
  eta = 0.3, gamma = 0.001,
  colsample_bytree = 0.75,
  min_child_weight = 1,
  subsample = 1
)

train_control <- caret::trainControl(
  method = "none",
  search = "grid",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
  allowParallel = TRUE
)

model <- caret::train(fatal_od_case ~ .,
  data = train.set,
  trControl = train_control,
  method = "xgbTree",
  tuneGrid = grid_default,
  verbose = TRUE
)

pred <- predict(model, test.set)
confusionMatrix(pred, test.set$fatal_od_case, positive = "Yes")

### Confusion Matrix and Statistics
###
## Reference
## Prediction  No  Yes
##        No  5980  936
##        Yes 3061 1026
##
## Accuracy : 0.6367
## 95% CI : (0.6277, 0.6457)
## No Information Rate : 0.8217
## P-Value [Acc > NIR] : 1
##
## Kappa : 0.1295
##
## Mcnemar's Test P-Value : <2e-16
##
## Sensitivity : 0.52294
## Specificity : 0.66143
## Pos Pred Value : 0.25104
## Neg Pred Value : 0.86466
## Prevalence : 0.17832
## Detection Rate : 0.09325
## Detection Prevalence : 0.37144
## Balanced Accuracy : 0.59218
##
## 'Positive' Class : Yes
##
Validating all the models

bothsample_rose_train_1 <- read_csv("U:/Andy's Thesis/MLoverdose/bothsample_rose_train_1.csv", show_col_types = FALSE)
datasettest <- read_csv("U:/Andy's Thesis/MLoverdose/test.set.csv", show_col_types = FALSE)

bothsample_rose_train_1 %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))
datasettest %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))

train.set <- bothsample_rose_train_1
test.setog <- datasettest
dim(train.set)
## [1] 25676  50
dim(test.setog)
## [1] 11003  50
nrow(train.set)
train.set = train.set %>% mutate_if(is.character, as.factor)

test.setog = test.setog %>% mutate_if(is.character, as.factor)

class(train.set$fatal_od_case)

## [1] "factor"

class(test.setog$fatal_od_case)

## [1] "factor"

table(train.set$fatal_od_case)

##
## No   Yes
## 12876 12800

#Xgboost

grid_default <- expand.grid(
  nrounds = 300,
  max_depth = 5,
  eta = 0.3, gamma = 0.001,
  colsample_bytree = 0.75,
  min_child_weight = 1,
  subsample = 1
)

train_control <- caret::trainControl(
  method = "cv",
  number = 10,
  search = "grid",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
  allowParallel = TRUE
)

class(train.set$fatal_od_case)

## [1] "factor"

XGBoostmodel = caret::train(fatal_od_case ~ ..,
  data = train.set,
  trControl = train_control,
  method = "xgbTree",
  tuneGrid = grid_default,
  verbose = TRUE
)
confusionMatrix(XGBmodel)

## Cross-Validated (10 fold) Confusion Matrix
## (entries are percentual average cell counts across resamples)
##
##       Reference
## Prediction  No  Yes
##        No  34.0 12.1
##        Yes 16.1 37.8
##
## Accuracy (average) : 0.718

confusionMatrix(XGBmodel$pred, XGBmodel$pred, XGBmodel$pred, XGBmodel$obs)

## Confusion Matrix and Statistics
##
##       Reference
## Prediction  No  Yes
##        No  8734 3098
##        Yes 4142 9702
##
##                Accuracy : 0.718
##                  95% CI : (0.7125, 0.7235)
##     No Information Rate : 0.5015
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4362
##
## McNemar's Test P-Value : < 2.2e-16
##
## Sensitivity : 0.6783
## Specificity : 0.7580
## Pos Pred Value : 0.7382
## Neg Pred Value : 0.7008
## Prevalence : 0.5015
## Detection Rate : 0.3402
## Detection Prevalence : 0.4608
## Balanced Accuracy : 0.7181
##
## 'Positive' Class : No
##
##Random Forest

grid_default <- expand.grid(
  .mtry = c(1:7),
  .splitrule="gini",
  .min.node.size=c(1)
)

train_control <- caret::trainControl(
method = "cv",
number = 10,
search = "grid",
summaryFunction = twoClassSummary,
returnResamp = "none",
classProbs = TRUE,
savePredictions = TRUE,
verboseIter = FALSE,
allowParallel = TRUE)
rf1 = caret::train(fatal_od_case ~ ,
data = train.set,
trControl = train_control,
method = "ranger",
tuneGrid = grid_default,
verbose = TRUE)
print(rf1)
## Random Forest
##
## 25676 samples
##    49 predictor
##     2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 23109, 23108, 23108, 23108, 23109, 23109, ...
## Resampling results across tuning parameters:
##
##   mtry  ROC        Sens       Spec
##   1     0.6776111  0.6213094  0.6362500
##   2     0.7614541  0.6752852  0.7035156
##   3     0.8207642  0.7141957  0.7685156
##   4     0.8576163  0.7335343  0.8053125
##   5     0.8791872  0.7464265  0.8309375
##   6     0.8929712  0.7560570  0.8478125
##   7     0.9022902  0.7667755  0.8603125
##
## Tuning parameter 'splitrule' was held constant at a value of gini
## Tuning parameter 'min.node.size' was held constant at a value of 1
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were mtry = 7, splitrule = gini
## and min.node.size = 1.
confusionMatrix(rf1)
## Cross-Validated (10 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
## Reference
## Prediction  No  Yes
##        No  38.5  7.0
##        Yes 11.7 42.9
##
## Accuracy (average) : 0.8134

```r
classificationMatrix(pred, obs)
```

## Confusion Matrix and Statistics
##
## Reference
## Prediction  No  Yes
##        No  64555 19806
##        Yes 25577 69794
##
## Accuracy : 0.7475
## 95% CI : (0.7455, 0.7495)
## No Information Rate : 0.5015
## P-Value [Acc > NIR] : < 2.2e-16
##
## Kappa : 0.4951
##
## McNemar's Test P-Value : < 2.2e-16
##
## Sensitivity : 0.7162
## Specificity : 0.7790
## Pos Pred Value : 0.7652
## Neg Pred Value : 0.7318
## Prevalence : 0.5015
## Detection Rate : 0.3592
## Detection Prevalence : 0.4694
## Balanced Accuracy : 0.7476
##
## 'Positive' Class : No
##

#Adabag

grid_default <- expand.grid(
  mfinal = 300,
  maxdepth = 5
)

train_control <- caret::trainControl(
  method = "cv",
  number = 10,
  search = "grid",
)
summaryFunction = twoClassSummary,
returnResamp = "none",
classProbs = TRUE,
savePredictions = TRUE,
verboseIter = FALSE,
allowParallel = TRUE
)

ada = caret::train(fatal_od_case ~ .,
data = train.set,
trControl = train_control,
method = "AdaBag",
tuneGrid = grid_default,
verbose = TRUE
)

print(ada)

## Bagged AdaBoost
##
## 25676 samples
##  49 predictor
##  2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 23108, 23109, 23109, 23109, 23108, 23108, ...
## Resampling results:
##
##   ROC        Sens       Spec
##   0.6446752  0.6470155  0.5594531
##
## Tuning parameter 'mfinal' was held constant at a value of 300
## Tuning parameter 'maxdepth' was held constant at a value of 5

confusionMatrix(ada)

## Cross-Validated (10 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
## Reference  No  Yes
## Prediction  No 32.4 22.0
##             Yes 17.7 27.9
##
## Accuracy (average) : 0.6034

confusionMatrix(ada$pred, ada$pred, obs)
## Confusion Matrix and Statistics

## Prediction  No  Yes
##        No  8331  5639
##        Yes  4545  7161

## Accuracy : 0.6034
## 95% CI : (0.5974, 0.6094)
## No Information Rate : 0.5015
## P-Value [Acc > NIR] : < 2.2e-16
## Kappa : 0.2065

## McNemar's Test P-Value : < 2.2e-16

## Sensitivity : 0.6470
## Specificity : 0.5595
## Pos Pred Value : 0.5963
## Neg Pred Value : 0.6117
## Prevalence : 0.5015
## Detection Rate : 0.3245
## Detection Prevalence : 0.5441
## Balanced Accuracy : 0.6032

'Positive' Class : No

# SVM

```r
train_control <- caret::trainControl(
  method = "cv",
  number = 10,
  search = "grid",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
  allowParallel = TRUE
)

svmmodel <- caret::train(fatal_od_case ~ .,
  data = train.set,
  trControl = train_control,
  method = "svmLinear",
  verbose = TRUE,
  tuneLength = 10
)
```

## line search fails -0.6560837 0.05802975 3.141064e-05 -1.433112e-05 -1.054484e-08 3.596e-09 -3.827549e-13line search fails -0.6452092 0.07072446 1.959825e-05 -1.057306e-05 -6.544206e-09 2.683846e-
## Support Vector Machines with Linear Kernel

## 25676 samples
## 49 predictor
## 2 classes: 'No', 'Yes'

## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 23109, 23109, 23108, 23108, 23108, 23109, ...
## Resampling results:

<table>
<thead>
<tr>
<th>ROC</th>
<th>Sens</th>
<th>Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6430621</td>
<td>0.5585873</td>
<td>0.6286458</td>
</tr>
</tbody>
</table>

## Tuning parameter 'C' was held constant at a value of 1

## Cross-Validated (10 fold) Confusion Matrix

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>28.0</td>
<td>18.5</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>22.1</td>
<td>31.3</td>
<td></td>
</tr>
</tbody>
</table>
## Accuracy (average) : 0.5935

**confusionMatrix**(svmmodel$pred, svmmodel$pred, svmmodel$obs)

## Confusion Matrix and Statistics
##
## Reference  
## Prediction   No  Yes  
##        No  4315 2852  
##        Yes 3410 4828  
##
## Accuracy : 0.5935
## 95% CI : (0.5857, 0.6013)
## No Information Rate : 0.5015
## P-Value [Acc > NIR] : < 2.2e-16
##
## Kappa : 0.1872
##
## Mcnemar's Test P-Value : 1.939e-12
##
## Sensitivity : 0.5586
## Specificity : 0.6286
## Pos Pred Value : 0.6021
## Neg Pred Value : 0.5861
## Prevalence : 0.5015
## Detection Rate : 0.2801
## Detection Prevalence : 0.4652
## Balanced Accuracy : 0.5936
##
## 'Positive' Class : No
##

#Logistic Regression

gird_default <- **expand.grid**(parameter=c(0.001, 0.01, 0.1, 1, 10, 100, 1000))

```r
train_control <- caret::**trainControl**(method = "cv",  
number = 10,  
search = "grid",  
summaryFunction = twoClassSummary,  
returnResamp = "none",  
classProbs = TRUE,  
savePredictions = TRUE,  
verboseIter = FALSE,  
allowParallel = TRUE)
```

LogitModel <- **train**(fatal_od_case ~ .,  
data = train.set,
method = "glm",
family = "binomial",
trControl = train_control,
tuneGrid = grid_default
)

print(LogitModel)

## Generalized Linear Model
## 25676 samples
## 49 predictor
## 2 classes: 'No', 'Yes'
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 23108, 23108, 23108, 23109, 23109, 23108, ...
## Resampling results across tuning parameters:
##
##   ROC        Sens     Spec      parameter
## 0.6461732  0.57533  0.636875  0.001
## 0.6461732  0.57533  0.636875  0.001
## 0.6461732  0.57533  0.636875  0.001
## 0.6461732  0.57533  0.636875  0.001
## 0.6461732  0.57533  0.636875  0.001
## 0.6461732  0.57533  0.636875  0.001
## 0.6461732  0.57533  0.636875  0.001

## ROC was used to select the optimal model using the largest value.
## The final value used for the model was parameter = 0.001.

confusionMatrix(LogitModel)

## Cross-Validated (10 fold) Confusion Matrix
## (entries are percentual average cell counts across resamples)
##
## Reference
## Prediction No Yes
## No 28.9 18.1
## Yes 21.3 31.7
##
## Accuracy (average) : 0.606

confusionMatrix(LogitModel$pred$pred, LogitModel$pred$obs)

## Confusion Matrix and Statistics
##
## Reference
## Prediction No Yes
## No 51856 32536
## Yes 38276 57064
##
## Accuracy : 0.606
## 95% CI : (0.6037, 0.6083)
## No Information Rate : 0.5015
## P-Value [Acc > NIR] : < 2.2e-16
##
## Kappa : 0.2122
##
## McNemar's Test P-Value : < 2.2e-16
##
## Sensitivity : 0.5753
## Specificity : 0.6369
## Pos Pred Value : 0.6145
## Neg Pred Value : 0.5985
## Prevalence : 0.5015
## Detection Rate : 0.2885
## Detection Prevalence : 0.4695
## Balanced Accuracy : 0.6061
##
## 'Positive' Class : No
##
## GBM

train_control <- caret::trainControl(
  method = "cv",
  number = 10,
  search = "grid",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
  allowParallel = TRUE
)

gbm <- train(fatal_od_case ~ .,
  data = train.set,
  method = "gbm",
  trControl = train_control,
  # tuneGrid = grid_default
)
```r
##
## 8  1.3606    nan  0.1000  0.0010
## 9  1.3587    nan  0.1000  0.0009
## 10 1.3570     nan  0.1000  0.0008
## 20 1.3452     nan  0.1000  0.0004
## 40 1.3350     nan  0.1000  0.0001
## 60 1.3301     nan  0.1000 -0.0001
## 80 1.3267     nan  0.1000  0.0000
##100 1.3239     nan  0.1000  0.0000
##120 1.3221     nan  0.1000  0.0000
##140 1.3204     nan  0.1000 -0.0000
##150 1.3197     nan  0.1000 -0.0000
##
print(gbm)
##
## Stochastic Gradient Boosting
##
## 25676 samples
## 49 predictor
## 2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 23109, 23108, 23108, 23108, 23109, 23109, ...
## Resampling results across tuning parameters:
##
## interaction.depth  n.trees  ROC       Sens     Spec
##    1       50  0.6344299  0.5729264  0.6201562
##    1     100  0.6418795  0.5639190  0.6355469
##    1    150  0.6444275  0.5697437  0.6338281
##    2       50  0.6468666  0.5800734  0.6325000
##    2     100  0.6575708  0.5892373  0.6369531
##    2    150  0.6626439  0.5937418  0.6414062
##    3       50  0.6548704  0.5992561  0.6246875
##    3     100  0.6548704  0.5992561  0.6246875
##    3    150  0.6726637  0.6049242  0.6457812
##
## Tuning parameter 'shrinkage' was held constant at a value of 0.1
##
## Tuning parameter 'n.minobsinnode' was held constant at a value of 10
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were n.trees = 150, interaction.depth =
## 3, shrinkage = 0.1 and n.minobsinnode = 10.

confusionMatrix(gbm)
##
## Cross-Validated (10 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
## Reference
```

## Prediction  No  Yes
##        No  30.3 17.7
##        Yes 19.8 32.2
##
## Accuracy (average) : 0.6253

```r
confusionMatrix(gbmPred$pred, gbmPred$obs)
```

### Confusion Matrix and Statistics

### Reference

### Prediction  No  Yes
###        No  67917 42132
###        Yes 47967 73068
###
### Accuracy : 0.6101
### 95% CI : (0.6081, 0.6121)
### No Information Rate : 0.5015
### P-Value [Acc > NIR] : < 2.2e-16
###
### Kappa : 0.2203
###
### McNemar's Test P-Value : < 2.2e-16
###
### Sensitivity : 0.5861
### Specificity : 0.6343
### Pos Pred Value : 0.6172
### Neg Pred Value : 0.6037
### Prevalence : 0.5015
### Detection Rate : 0.2939
### Detection Prevalence : 0.4762
### Balanced Accuracy : 0.6102
###
### 'Positive' Class : No
###

# Naive Bayes

```r
train_control <- caret::trainControl(
  method = "cv",
  number = 10,
  search = "grid",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
  allowParallel = TRUE
)

nBmodel <- train(fatal_od_case ~ .,
                 data = train.set,
```
method = "nb",
          trControl = train_control
  )

confusionMatrix(nBmodel)

## Cross-Validated (10 fold) Confusion Matrix
## (entries are percentual average cell counts across resamples)
##
##        Reference
## Prediction   No  Yes
##        No  27.1 17.9
##        Yes 23.1 32.0
##
## Accuracy (average) : 0.5904

confusionMatrix(nBmodel$pred, nBmodel$obs)

## Confusion Matrix and Statistics
##
##        Reference
## Prediction   No  Yes
##        No  14459  9764
##        Yes 11293 15836
##
## Accuracy : 0.5899
## 95% CI : (0.5857, 0.5942)
## No Information Rate : 0.5015
## P-Value [Acc > NIR] : < 2.2e-16
##
## Kappa : 0.18
##
## Mcnemar's Test P-Value : < 2.2e-16
##
## Sensitivity : 0.5615
## Specificity : 0.6186
## Pos Pred Value : 0.5969
## Neg Pred Value : 0.5837
## Prevalence : 0.5015
## Detection Rate : 0.2816
## Detection Prevalence : 0.4717
## Balanced Accuracy : 0.5900
##
## 'Positive' Class : No
##
## #KNN

grid_default <- expand.grid(k = 1)
train_control <- caret::trainControl(
  method = "cv",
  number = 10,
  search = "grid",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
  allowParallel = TRUE
)

cnnmodel <- train(fatal_od_case ~ .,
  data = train.set,
  method = "knn",
  trControl = train_control,
  tuneGrid = grid_default)

cnfusionMatrix(cnnmodel)

## Cross-Validated (10 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##           Reference
## Prediction  No  Yes
##        No  35.6  5.1
##        Yes 14.5 44.8
##
##  Accuracy (average) : 0.8039

cnfusionMatrix(cnnmodel$pred, cnnmodel$pred)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No  Yes
##        No  9148 1308
##        Yes 3728 11492
##
##  Accuracy : 0.8039
##  95% CI : (0.799, 0.8087)
## No Information Rate : 0.5015
## P-Value [Acc > NIR] : < 2.2e-16
##
##  Kappa : 0.6079
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##  Sensitivity : 0.7105
##  Specificity : 0.8978
##  Pos Pred Value : 0.8749
## Neg Pred Value : 0.7551
## Prevalence : 0.5015
## Detection Rate : 0.3563
## Detection Prevalence : 0.4072
## Balanced Accuracy : 0.8041
## 'Positive' Class : No
##
# Testing for the best performing model

# Xgboost
predt <- predict(XGBoostmodel, test.setog)
confusionMatrix(predt, test.setog$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference
## Prediction  No  Yes
## No  5968  939
## Yes  3073 1023
##
## Accuracy : 0.6354
## 95% CI : (0.6263, 0.6444)
## No Information Rate : 0.8217
## P-Value [Acc > NIR] : 1
##
## Kappa : 0.1273
##
## McNemar's Test P-Value : <2e-16
##
## Sensitivity : 0.52141
## Specificity : 0.66010
## Pos Pred Value : 0.24976
## Neg Pred Value : 0.86405
## Prevalence : 0.17832
## Detection Rate : 0.09297
## Detection Prevalence : 0.37226
## Balanced Accuracy : 0.59076
##
## 'Positive' Class : Yes
##
prob <- predict(XGBoostmodel, test.setog, type = "prob")
results.roc <- roc(test.setog$fatal_od_case, prob$Yes)
auc(results.roc)

## Area under the curve: 0.6212

plot(results.roc, print.thres = "best", printe.thres.best.method = "closest.topleft")
```
results.coords <- coords(results.roc,"best", best.method = "closest.topleft")
print(results.coords)
## threshold specificity sensitivity
## 1 0.4573449   0.5918593    0.588685

#RF
rf1t <- predict(rf1, test.setog)
confusionMatrix(rf1t, test.setog$fatal_od_case, positive = "Yes")
## Confusion Matrix and Statistics
##                       Reference
## Prediction   No  Yes
##        No  6535 1070
##        Yes 2506  892
##
##                Accuracy : 0.675
##                  95% CI : (0.6662, 0.6837)
##     No Information Rate : 0.8217
##     P-Value [Acc > NIR] : 1
##
##                Kappa : 0.1379
##
##    Mcnemar's Test P-Value : <2e-16
##
##    Sensitivity : 0.45464
##    Specificity : 0.72282
```
##  Pos Pred Value : 0.26251  
##  Neg Pred Value : 0.85930  
##  Prevalence : 0.17832  
##  Detection Rate : 0.08107  
##  Detection Prevalence : 0.30882  
##  Balanced Accuracy : 0.58873  
##  'Positive' Class : Yes
##

```
prob <- predict(rf1, test.setog, type = "prob")
results.roc1 <- roc(test.setog$fatal_od_case, prob$Yes)
auc(results.roc1)
## Area under the curve: 0.6352
```

```
plot(results.roc1, print.thres = "best", printe.thres.best.method = "closest.topleft")
```

```
results.coords <- coords(results.roc1,"best", best.method = "closest.topleft")
print(results.coords)
##  threshold specificity sensitivity
## 1  0.4231978   0.5991594    0.598369
```

```
# AdaBag
adat <- predict(ada, test.setog)
confusionMatrix(adat, test.setog$fatal_od_case, positive = "Yes")
```
## Confusion Matrix and Statistics

### Reference
Prediction | No | Yes
---|---|---
No | 5734 | 869
Yes | 3036 | 1080

### Accuracy : 0.6357

95% CI : (0.6269, 0.645)

No Information Rate : 0.8217

P-Value [Acc > NIR] : 1

Kappa : 0.1433

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.55413
Specificity : 0.65382
Pos Pred Value : 0.26239
Neg Pred Value : 0.86839
Prevalence : 0.17832
Detection Rate : 0.09879
Detection Prevalence : 0.38326
Balanced Accuracy : 0.60391

'Positive' Class : Yes

```r
prob <- predict(ada, test.setog, type = "prob")
results.roc2 <- roc(test.setog$fatal_od_case, prob$Yes)
auc(results.roc2)

# Area under the curve: 0.6424
```

plot(results.roc2, print.thres = "best", print.thres.best.method = "closest.topleft")
results.coords <- coords(results.roc2,"best", best.method = "closest.topleft")
print(results.coords)

## threshold specificity sensitivity
## 1 0.4016667   0.6218339     0.59684

#Support Vector Machine

svmpt <- predict(svmmodel, test.setog)
confusionMatrix(factor(svmpt),factor(test.setog$fatal_od_case))

## Confusion Matrix and Statistics
## Reference
## Prediction  No  Yes
##    No  5396  724
##    Yes 3916 1251
##
## Accuracy : 0.5889
## 95% CI : (0.5798, 0.598)
## No Information Rate : 0.825
## P-Value [Acc > NIR] : 1
##
## Kappa : 0.1301
##
## McNemar's Test P-Value : <2e-16
##
Sensitivity : 0.5795
Specificity : 0.6334
Pos Pred Value : 0.8817
Neg Pred Value : 0.2421
Prevalence : 0.8250
Detection Rate : 0.4781
Detection Prevalence : 0.5422
Balanced Accuracy : 0.6064

'Positive' Class : No

```
prob <- predict(svmmodel, test.setog, type = "prob")
results.roc3 <- roc(test.setog$fatal_od_case, prob$Yes)
auc(results.roc3)

## Area under the curve: 0.6542
plot(results.roc3, print.thres = "best", print.e.thres.best.method = "closest.topleft")
```

```
results.coords <- coords(results.roc3, "best", best.method = "closest.topleft")
print(results.coords)

## threshold specificity sensitivity
## 1 0.5282977 0.647122 0.5792405
```
# Logistic Regression

LogitModelt <- `predict`(LogitModel, newdata = test.setog)

confusionMatrix(LogitModelt, test.setog$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
### Reference
### Prediction  No  Yes
###     No  5368  740
###     Yes 3673 1222
###
### Accuracy : 0.5989
### 95% CI : (0.5897, 0.6081)
### No Information Rate : 0.8217
### P-Value [Acc > NIR] : 1
###
### Kappa : 0.1366
###
### McNemar's Test P-Value : <2e-16
###
### Sensitivity : 0.6228
### Specificity : 0.5937
### Pos Pred Value : 0.2496
### Neg Pred Value : 0.8788
### Prevalence : 0.1783
### Detection Rate : 0.1111
### Detection Prevalence : 0.4449
### Balanced Accuracy : 0.6083
###
### 'Positive' Class : Yes
###

prob <- `predict`(LogitModel, test.setog, type = "prob")

results.roc4 <- roc(test.setog$fatal_od_case, prob$Yes)

auc(results.roc4)

## Area under the curve: 0.6499

plot(results.roc4, print.thres = "best", printe.thres.best.method = "closest.topleft")
results.coords <- coords(results.roc4, "best", best.method = "closest.topleft")
print(results.coords)

## threshold specificity sensitivity
## 1 0.5077293 0.6107731 0.6075433

#GBM
gbmt <- predict(gbm, newdata = test.setog)
confusionMatrix(gbmt, test.setog$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference
## Prediction   No  Yes
##        No  5475  759
##        Yes 3566 1203
##
##                Accuracy : 0.6069
##                  95% CI : (0.5977, 0.6161)
##     No Information Rate : 0.8217
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1402
##
## McNemar's Test P-Value : <2e-16
##
## Sensitivity : 0.6131
## Specificity : 0.6056
##          Pos Pred Value : 0.2523
##          Neg Pred Value : 0.8782
##              Prevalence : 0.1783
##          Detection Rate : 0.1093
##    Detection Prevalence : 0.4334
##       Balanced Accuracy : 0.6094
##
##        'Positive' Class : Yes
##
prob <- predict(gbm, test.setog, type = "prob")
results.roc5 <- roc(test.setog$fatal_od_case, prob$Yes)
auc(results.roc5)

## Area under the curve: 0.6612
plot(results.roc5, print.thres = "best", printe.thres.best.method = "closest.topleft")

results.coords <- coords(results.roc5,"best", best.method = "closest.topleft")
print(results.coords)

##  threshold specificity sensitivity
## 1  0.4891762  0.5813516  0.648318

#Naive Bayes
Nbpt <- predict(nBmodel, newdata = test.setog)
confusionMatrix(Nbpt, test.setog$fatal_od_case, positive = "Yes")
## Confusion Matrix and Statistics

### Reference
- No 5073 767
- Yes 3968 1195

### Accuracy: 0.5697
- 95% CI: (0.5603, 0.5789)
- No Information Rate: 0.8217
- P-Value [Acc > NIR]: 1

### Kappa: 0.1039

### McNemar's Test P-Value: <2e-16

### Sensitivity: 0.6091
### Specificity: 0.5611
### Pos Pred Value: 0.2315
### Neg Pred Value: 0.8687
### Prevalence: 0.1783
### Detection Rate: 0.1086
### Detection Prevalence: 0.4692
### Balanced Accuracy: 0.5851

### 'Positive' Class: Yes

```r
prob <- predict(nBmodel, test.setog, type = "prob")
results.roc6 <- roc(test.setog$fatal_od_case, prob$Yes)
auc(results.roc6)
```

### Area under the curve: 0.6268

```r
plot(results.roc6, print.thres = "best", print.thres.best.method = "closest.topleft")
```
results.coords <- \texttt{coords}(results.roc6, \texttt{"best"}, \texttt{best.method = \"closest.topleft\")
print(results.coords)

## threshold specificity sensitivity
## 1 0.5692031 0.6242672 0.558104

\textbf{#KNN}

Knnpt <- \texttt{predict}(knnmodel, \texttt{newdata = test.setog})
\texttt{confusionMatrix}(Knnpt, test.setog$fatal_od_case, \texttt{positive = \"Yes\")

## Confusion Matrix and Statistics
##
## Reference
## Prediction  No Yes
##   No  5530  986
##   Yes 3511  976
##
## Accuracy: 0.5913
## 95\% CI: (0.582, 0.6005)
## No Information Rate: 0.8217
## P-Value [Acc > NIR]: 1
##
## Kappa: 0.0726
##
## Mcnemar's Test P-Value: <2e-16
##
## Sensitivity: 0.4975
## Specificity: 0.6117
##          Pos Pred Value : 0.2175
##          Neg Pred Value : 0.8487
##          Prevalence : 0.1783
##          Detection Rate : 0.0887
##    Detection Prevalence : 0.4078
##          Balanced Accuracy : 0.5546
##        'Positive' Class : Yes
##

```r
prob <- predict(knnmodel, test.setog, type = "prob")
results.roc7 <- roc(test.setogfatal_od_case, prob$Yes)
auc(results.roc7)
## Area under the curve: 0.5602
plot(results.roc7, print.thres = "best", printe.thres.best.method = "closest.topleft")
```

```r
results.coords <- coords(results.roc7, "best", best.method = "closest.topleft")
print(results.coords)
##  threshold specificity sensitivity
##   1 0.4753834   0.5737197   0.5407747
```

ML deep learning with Keras

```r
library(keras)
library(tensorflow)
```
library(tidyverse)

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x purrr::%@%()         masks rlang::%@%()
## x dplyr::filter()     masks stats::filter()
## x purrr::flatten()    masks rlang::flatten()
## x purrr::flatten_chr() masks rlang::flatten_chr()
## x purrr::flattendbl() masks rlang::flattendbl()
## x purrr::flatten_int() masks rlang::flatten_int()
## x purrr::flatten_lgl() masks rlang::flatten_lgl()
## x purrr::flatten_raw() masks rlang::flatten_raw()
## x purrr::invoke()     masks rlang::invoke()
## x dplyr::lag()        masks stats::lag()
## x purrr::splice()     masks rlang::splice()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(dplyr)
library(gapminder)
library(caret)

## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
## lift
##
## The following object is masked from 'package:tensorflow':
##
## train

library(xgboost)

##
## Attaching package: 'xgboost'
##
## The following object is masked from 'package:dplyr':
##
## slice

library(forcats)
library(mice)
library(vctrs)

library(naniar)
library(yardstick)

## For binary classification, the first factor level is assumed to be the event.
## Use the argument `event_level = "second"` to alter this as needed.

library(forcats)

datasettrain <- read_csv("U:/Andy's Thesis/MLoverdose/bothsample_rose_train_1.csv", show_col_types = FALSE)

datasettest <- read_csv("U:/Andy's Thesis/MLoverdose/test.set.csv", show_col_types = FALSE)
train.set <- datasettrain
test.set <- datasettest

dim(train.set)
## [1] 25676 50
dim(test.set)
## [1] 11003 50

nrow(train.set)
## [1] 25676

train.set = train.set %>% mutate_if(is.character, as.factor)
test.set = test.set %>% mutate_if(is.character, as.factor)

class(train.set$fatal_od_case)
## [1] "numeric"
class(test.set$fatal_od_case)
## [1] "numeric"
table(train.set$fatal_od_case)
##
##    0   1
## 12876 12800

table(test.set$fatal_od_case)
##
##    0   1
## 9041 1962

x_train_tbl <- train.set %>% select(-fatal_od_case)
x_test_tbl <- test.set %>% select(-fatal_od_case)

y_train_vec <- train.set %>% select(fatal_od_case)

y_test_vec <- test.set %>% select(fatal_od_case)
glimpse(x_train_tbl)
model_keras <- keras_model_sequential()

model_keras %>%
  layer_dense(
    units = 49,
    activation = "relu",
    input_shape = ncol(x_train_tbl) >>
  ) %>%
  layer_dropout(rate = 0.001) %>%
  layer_dense(
    units = 49,
    activation = "relu"
  ) %>%
  layer_dropout(rate = 0.001) %>%
  layer_dense(
    units = 49,
    activation = "relu"
  ) %>%
  layer_dropout(rate = 0.001) %>%
  layer_dense(
    units = 49,
    activation = "relu"
  ) %>%
  layer_dropout(rate = 0.001) %>%
  layer_dense(
    units = 1,
    activation = "sigmoid"
  ) %>%
  compile(  
    optimizer = 'adam',
    loss = 'binary_crossentropy',
    metrics = c('accuracy')
  )

model_keras

## Model
##
##  Layer (type)                        Output Shape                    Param #  
##  =================================================================
##  dense_5 (Dense)                     (None, 49)                      2450   
##  dropout_4 (Dropout)                 (None, 49)                      0      
##  dense_4 (Dense)                     (None, 49)                      2450   
##  dropout_3 (Dropout)                 (None, 49)                      0      

## dense_3 (Dense) (None, 49) 2450
## dropout_2 (Dropout) (None, 49) 0
## dense_2 (Dense) (None, 49) 2450
## dropout_1 (Dropout) (None, 49) 0
## dense_1 (Dense) (None, 49) 2450
## dropout (Dropout) (None, 49) 0
## dense (Dense) (None, 1) 50

---

## Total params: 12,300
## Trainable params: 12,300
## Non-trainable params: 0

history <- fit(
  object = model keras,
  x = as.matrix(x_train_tbl),
  y = as.matrix(y_train_vec),
  batch_size = 300,
  epochs = 1000,
  validation_split = 0.3
)

print(history)

## Trained on 17,973 samples (batch_size=300, epochs=1,000)
## Final epoch (plot to see history):
##     loss: 0.1669
##      acc: 0.9164
## val_loss: 1.656
##  val_acc: 0.6037

plot(history)

## `geom_smooth()` using formula = 'y ~ x'
yhat_keras_class_vec <- predict_classes(object = model_keras, x = as.matrix(x_test_tbl)) %>%
  as.vector()

yhat_keras_prob_vec <- predict_proba(object = model_keras, x = as.matrix(x_test_tbl)) %>%
  as.vector()

estimates_tbl <- y_test_vec %>%
  add_column(
    estimate = as.factor(yhat_keras_class_vec),
    class_prob = yhat_keras_prob_vec
  )

estimates_tbl

# A tibble: 11,003 x 3
#  fatal_od_case estimate class_prob
#   <dbl> <fct>      <dbl>
# 1     0 0          0
# 2     0 0          0.000147
# 3     0 0          0.000000209
# 4     0 0          0
# 5     0 0          0.0166
# 6     0 0          0.000725
# 7     0 0          0.0000177
# 8     0 0          0
# 9     0 0          0.00000864
estimates_tbl = estimates_tbl %>% mutate_at(vars(fatal_od_case), list(factor))

class(estimates_tbl$fatal_od_case)
## [1] "factor"

levels(estimates_tbl$fatal_od_case)
## [1] "0" "1"

levels(estimates_tbl$estimate)
## [1] "0" "1"

options(yardstick.event_first = FALSE)

estimates_tbl %>% conf_mat(fatal_od_case, estimate)
##           Truth
## Prediction    0    1
##          0 7338 1467
##          1 1703  495

estimates_tbl %>% metrics(fatal_od_case, estimate)
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary        0.712
## 2 kap      binary        0.0611

estimates_tbl %>% roc_auc(fatal_od_case, class_prob)
## Warning: The `yardstick.event_first` option has been deprecated as of yardstick 0.0.7 and will be completely ignored in a future version.
## Instead, set the following argument directly in the metric function:
## `options(yardstick.event_first = TRUE)` -> `event_level = 'first'` (the default)
## `options(yardstick.event_first = FALSE)` -> `event_level = 'second'`
## This warning is displayed once per session.

## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.576

ML modeling with Tidy models

library(keras)
library(tensorflow)
library(rlang)
library(tidyverse)
library(dplyr)
library(gapminder)
library(xgboost)
library(forcats)
library(mice)
library(vctrs)
library(naniar)
library(yardstick)
library(forcats)
library(Boruta)
library(ROSE)
library(imbalance)
library(rnn)
library(ROCR)
library(SHAPforxgboost)
library(randomForest)
library(mlbench)
library(ggplot2)
library(data.table)
library(here)
library(tidymodels)
library(parsnip)
library(caret)
library(kknn)
library(yardstick)
library(tune)

### Testing the original sample

datasettrain <- read_csv("U:/Andy's Thesis/MLoverdose/train.set.csv", show_col_types = FALSE)
datasettest <- read_csv("U:/Andy's Thesis/MLoverdose/test.set.csv", show_col_types = FALSE)
datasettrain <- datasettrain %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))
datasettest <- datasettest %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))
train.set <- datasettrain
test.set <- datasettest
dim(train.set)
## [1] 25676  50
dim(test.set)
## [1] 11003  50
nrow(train.set)
## [1] 25676

```r
train.set = train.set %>% mutate_if(is.character, as.factor)
test.set = test.set %>% mutate_if(is.character, as.factor)

class(train.set$fatal_od_case)
## [1] "factor"
class(test.set$fatal_od_case)
## [1] "factor"
table(train.set$fatal_od_case)
##
##   No   Yes
## 21212  4464

#XGBoost

xgb_spec <- boost_tree(
  trees = 300) %>%
  set_engine("xgboost") %>%
  set_mode("classification")

overdose_rec <-
  recipe(fatal_od_case ~., data = train.set)

xgbworkflow <- workflow() %>%
  add_model(xgb_spec) %>%
  add_recipe(overdose_rec)

xgb_fit <- xgbworkflow %>%
  fit(data = train.set)

class_pred <- predict(xgb_fit, test.set)
prob_pred <- predict(xgb_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("XGB_Class", "LR_DeathProb", "R_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)
caret::confusionMatrix(ODD_preds$XGB_Class, ODD_preds$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference  Prediction
## No          8789 1848
## Yes         252  114
# Random Forest

```r
rf_mod <-
  rand_forest(trees = 500) %>%
  set_engine("ranger") %>%
  set_mode("classification")

rf_fit <-
  rf_mod %>%
  fit(fatal_od_case ~ ., data = train.set)

class_pred <- predict(rf_fit, test.set)
prob_pred <- predict(rf_fit, test.set, type = "prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("RF_Class", "RF_DeathProb", "RF_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)

table(ODD_preds$fatal_od_case)

# Confusion Matrix and Statistics
# Reference

caret::confusionMatrix(ODD_preds$RF_Class, ODD_preds$fatal_od_case, positive = "Yes")
```
## Prediction   No  Yes
##        No  9014 1929
##        Yes   27   33
##                Accuracy : 0.8222
##                  95% CI : (0.815, 0.8293)
##     No Information Rate : 0.8217
##     P-Value [Acc > NIR] : 0.4466
##                Kappa : 0.0223
## Mcnemar's Test P-Value : <2e-16
##                Sensitivity : 0.016820
##                Specificity : 0.997014
##              Pos Pred Value : 0.550000
##              Neg Pred Value : 0.823723
##               Prevalence : 0.178315
##          Detection Rate : 0.002999
##    Detection Prevalence : 0.005453
##       Balanced Accuracy : 0.506917
##        'Positive' Class : Yes

### Testing Different Samples

```r
# Undersample 3:1
undersample_rose_train_1 <- read_csv("U:/Andy's Thesis/MLoverdose/undersample_rose_train_1.csv", show_col_types = FALSE)

undersample_rose_train_1 <- undersample_rose_train_1 %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))

train.set <- undersample_rose_train_1

train.set = train.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 16503  50

dim(test.set)
## [1] 11003  50

nrow(train.set)
## [1] 16503

#XGBoost
xgb_spec <- \texttt{boost.tree}(
  \texttt{trees = 300}) \texttt{set_engine}("\texttt{xgboost}) \texttt{set_mode}"("classification")

overdose_rec <- \texttt{recipe}(\texttt{fatal_od_case} \sim ., \texttt{data = train.set})

xgbworkflow <- \texttt{workflow()} \texttt{add_model}(xgb_spec) \texttt{add_recipe}(overdose_rec)

xgb_fit <- xgbworkflow \texttt{fit}(\texttt{data = train.set})

class_pred <- \texttt{predict}(xgb_fit, test.set)
prob_pred <- \texttt{predict}(xgb_fit, test.set, \texttt{type = "prob"})
predictions <- \texttt{data.frame}(\texttt{class_pred, prob_pred}) \texttt{setNames}(\texttt{c("XGB\_Class", "LR\_DeathProb", "R\_NotDeathProb")})

ODD_preds <- test.set \texttt{bind_cols}(predictions)
caret::\texttt{confusionMatrix}(ODD_preds$XGB\_Class, ODD_preds$fatal_od_case, \texttt{positive = "Yes")}

## Confusion Matrix and Statistics
##
## Prediction  No  Yes
## No   8203 1626
## Yes  838 336
##
## Accuracy : 0.7761
## 95% CI : (0.7682, 0.7838)
## No Information Rate : 0.8217
## P-Value [Acc > NIR] : 1
##
## Kappa : 0.0932
##
## McNemar's Test P-Value : <2e-16
##
## Sensitivity : 0.17125
## Specificity : 0.90731
## Pos Pred Value : 0.28620
## Neg Pred Value : 0.83457
## Prevalence : 0.17832
## Detection Rate : 0.03054
## Detection Prevalence : 0.10670
## Balanced Accuracy : 0.53928
##
'Positive' Class : Yes

class(train.set$sfatal_od_case)
## [1] "factor"

class(test.set$sfatal_od_case)
## [1] "factor"

class(class_pred$s.pred_class)
## [1] "factor"
class_pred <- predict(xgb_fit, test.set)
caret::confusionMatrix(class_pred$s.pred_class, test.set$sfatal_od_case)

Confusion Matrix and Statistics

Reference
Prediction  No  Yes
No  8203 1626
Yes  838  336

Accuracy : 0.7761
95% CI : (0.7682, 0.7838)
No Information Rate : 0.8217
P-Value [Acc > NIR] : 1

Kappa : 0.0932

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.9073
Specificity : 0.1713
Pos Pred Value : 0.8346
Neg Pred Value : 0.2862
Prevalence : 0.8217
Detection Rate : 0.7455
Detection Prevalence : 0.8933
Balanced Accuracy : 0.5393

'Positive' Class : No

#Random Forest

rf_mod <-
  rand_forest(trees=500) %>%
  set_engine("ranger") %>%
  set_mode("classification")
rf_fit <-
  rf_mod %>%
  fit(fatal_od_case ~., data = train.set)

class_pred <- predict(rf_fit, test.set)
prob_pred <- predict(rf_fit, test.set, type = "prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("RF_Class", "RF_DeathProb", "RF_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)
caret::confusionMatrix(ODD_preds$RF_Class, ODD_preds$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
## Reference
## Prediction  No  Yes
##        No  8758 1804
##        Yes  283  158
##
##    Accuracy : 0.8103
##  95% CI : (0.8029, 0.8176)
## No Information Rate : 0.8217
## P-Value [Acc > NIR] : 0.999
##
##    Kappa : 0.0707
##
## McNemar's Test P-Value : <2e-16
##
##    Sensitivity : 0.08053
##    Specificity : 0.96870
##    Pos Pred Value : 0.35828
##    Neg Pred Value : 0.82920
##    Prevalence : 0.17832
##    Detection Rate : 0.01436
##    Detection Prevalence : 0.04008
##    Balanced Accuracy : 0.52461
##
## 'Positive' Class : Yes
##

class_pred <- predict(rf_fit, test.set)
caret::confusionMatrix(class_pred$pred_class, test.set$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
## Reference
## Prediction  No  Yes
##        No  8758 1804
##        Yes  283  158
##
##                Accuracy : 0.8103
##                  95% CI : (0.8029, 0.8176)
##     No Information Rate : 0.8217
##     P-Value [Acc > NIR] : 0.999
##
##                   Kappa : 0.0707
##
##  Mcnemar's Test P-Value : <2e-16
##
##                  Sensitivity : 0.08053
##                  Specificity : 0.96870
##                  Pos Pred Value : 0.35828
##                  Neg Pred Value : 0.82920
##                  Prevalence : 0.17832
##                  Detection Rate : 0.01436
##             Detection Prevalence : 0.04008
##               Balanced Accuracy : 0.52461
##
##        'Positive' Class : Yes
##
##Undersample 1:1

undersample_rose_train_2 <- read_csv("U:/Andy's Thesis/MLoverdose/undersample_rose_train_2.csv", show_col_types = FALSE)

undersample_rose_train_2 <- undersample_rose_train_2 %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))

train.set <- undersample_rose_train_2

train.set = train.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 8917   50

dim(test.set)
## [1] 11003   50

nrow(train.set)
## [1] 8917

#XGBoost

xgb_spec <- boost_tree(
  trees = 300) %>%
  set_engine("xgboost") %>%
  set_mode("classification")

overdose_rec <-
recipe(fatal_od_case ~., data = train.set)

xgbworkflow <- workflow()
  add_model(xgb_spec)
  add_recipe(overdose_rec)

xgb_fit <- xgbworkflow
fit(data = train.set)

class_pred <- predict(xgb_fit, test.set)
prob_pred <- predict(xgb_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred)
  setNames(c("XGB_Class", "LR_DeathProb", "R_NotDeathProb"))

ODD_preds <- test.set
  bind_cols(predictions)

caret::confusionMatrix(ODD_preds$XGB_Class, ODD_preds$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference
##        No  Yes
##  No  5215  785
##  Yes 3826 1177
##
## Accuracy : 0.5809
##  95% CI : (0.5716, 0.5902)
## No Information Rate : 0.8217
## P-Value [Acc > NIR] : 1
##
## Kappa : 0.11
##
## Mcnemar's Test P-Value : <2e-16
##
## Sensitivity : 0.5999
## Specificity : 0.5768
## Pos Pred Value : 0.2353
## Neg Pred Value : 0.8692
## Prevalence : 0.1783
## Detection Rate : 0.1070
## Detection Prevalence : 0.4547
## Balanced Accuracy : 0.5884
##
## 'Positive' Class : Yes
##

class_pred <- predict(xgb_fit, test.set)
caret::confusionMatrix(class_pred$pred_class, test.set$fatal_od_case, positive = "Yes")
## Confusion Matrix and Statistics

### Reference
### Prediction  No  Yes
### No  5215  785
### Yes 3826 1177
### Accuracy : 0.5809
### 95% CI : (0.5716, 0.5902)
### No Information Rate : 0.8217
### P-Value [Acc > NIR] : 1
### Kappa : 0.11
### McNemar's Test P-Value : <2e-16
### Sensitivity : 0.5999
### Specificity : 0.5768
### Pos Pred Value : 0.2353
### Neg Pred Value : 0.8692
### Prevalence : 0.1783
### Detection Rate : 0.1070
### Detection Prevalence : 0.4547
### Balanced Accuracy : 0.5884
### 'Positive' Class : Yes

### Random Forest

```r
#Random Forest

rf_mod <-
  rand_forest(trees=500) %>%
  set_engine("ranger") %>%
  set_mode("classification")

rf_fit <-
  rf_mod %>%
  fit(fatal_od_case ~., data = train.set)

class_pred <- predict(rf_fit, test.set)
prob_pred <- predict(rf_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("RF_Class", "RF_DeathProb", "RF_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)

caret::confusionMatrix(ODD_preds$RF_Class, ODD_preds$fatal_od_case, positive = "Yes")
```

## Confusion Matrix and Statistics

### Reference
## Prediction  No  Yes
##        No  5512  765
##        Yes 3529 1197
##
##                Accuracy : 0.6097
##                  95% CI : (0.6006, 0.6189)
##     No Information Rate : 0.8217
##     P-Value [Acc > NIR] : 1
##
##                Kappa : 0.1416
##
## Mcnemar's Test P-Value : <2e-16
##
##                Sensitivity : 0.6101
##                Specificity : 0.6097
##                Pos Pred Value : 0.2533
##                Neg Pred Value : 0.8781
##                Prevalence : 0.1783
##                Detection Rate : 0.1088
##        Detection Prevalence : 0.4295
##                Balanced Accuracy : 0.6099
##
##       'Positive' Class : Yes
##

class_pred <- predict(rf_fit, test.set)
caret::confusionMatrix(class_pred$pred_class, test.set$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
##                Reference
##                Prediction  No  Yes
##        No  5512  765
##        Yes 3529 1197
##
##                Accuracy : 0.6097
##                  95% CI : (0.6006, 0.6189)
##     No Information Rate : 0.8217
##     P-Value [Acc > NIR] : 1
##
##                Kappa : 0.1416
##
## Mcnemar's Test P-Value : <2e-16
##
##                Sensitivity : 0.6101
##                Specificity : 0.6097
##                Pos Pred Value : 0.2533
##                Neg Pred Value : 0.8781
##                Prevalence : 0.1783
##                Detection Rate : 0.1088
##        Detection Prevalence : 0.4295
## Balanced Accuracy : 0.6099
## 'Positive' Class : Yes
##
## Bothsample 1:1

```r
bothsample_rose_train_1 <- read_csv("U:/Andy's Thesis/MLoverdose/bothsample_rose_train_1.csv", show_col_types = FALSE)
bothsample_rose_train_1 <- bothsample_rose_train_1 %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))
train.set <- bothsample_rose_train_1

train.set = train.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 25676   50

dim(test.set)
## [1] 11003   50

nrow(train.set)
## [1] 25676

# XGBoost

xgb_spec <- boost_tree(trees = 300) %>%
  set_engine("xgboost") %>%
  set_mode("classification")

overdose_rec <-
  recipe(fatal_od_case ~ ., data = train.set)

xgbworkflow <- workflow() %>%
  add_model(xgb_spec) %>%
  add_recipe(overdose_rec)

xgb_fit <- xgbworkflow %>%
  fit(data = train.set)

class_pred <- predict(xgb_fit, test.set)
prob_pred <- predict(xgb_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("XGB_Class", "LR_DeathProb", "R_NotDeathProb"))
```
ODD_preds <- test.set %>%
  bind_cols(predictions)
caret::confusionMatrix(ODD_preds$XGB_Class, ODD_preds$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference
## Prediction  No  Yes
##        No  6117  972
##        Yes 2924  990
##
## Accuracy : 0.6459
## 95% CI : (0.6369, 0.6549)
## No Information Rate : 0.8217
## P-Value [Acc > NIR] : 1
##
## Kappa : 0.1304
##
## Mcnemar's Test P-Value : <2e-16
##
## Sensitivity : 0.50459
## Specificity : 0.67658
## Pos Pred Value : 0.25294
## Neg Pred Value : 0.86289
## Prevalence : 0.17832
## Detection Rate : 0.08998
## Detection Prevalence : 0.35572
## Balanced Accuracy : 0.59059
##
## 'Positive' Class : Yes
##
class_pred <- predict(xgb_fit, test.set)
caret::confusionMatrix(class_pred$.pred_class, test.set$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference
## Prediction  No  Yes
##        No  6117  972
##        Yes 2924  990
##
## Accuracy : 0.6459
## 95% CI : (0.6369, 0.6549)
## No Information Rate : 0.8217
## P-Value [Acc > NIR] : 1
##
## Kappa : 0.1304
##
## Mcnemar's Test P-Value : <2e-16
##
## Sensitivity : 0.5045
## Specificity : 0.6765
## Pos Pred Value : 0.2529
## Neg Pred Value : 0.8629
## Prevalence : 0.1783
## Detection Rate : 0.0899
## Detection Prevalence : 0.3557
## Balanced Accuracy : 0.5906

### 'Positive' Class : Yes

### Random Forest

```r
rf_mod <-
  rand_forest(trees=500) %>%
  set_engine("ranger") %>%
  set_mode("classification")
rf_fit <-
  rf_mod %>%
  fit(fatal_od_case ~., data = train.set)
class_pred <- predict(rf_fit, test.set)
prob_pred <- predict(rf_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("RF_Class", "RF_DeathProb", "RF_NotDeathProb"))
ODD_preds <- test.set %>%
  bind_cols(predictions)
caret::confusionMatrix(ODD_preds$RF_Class, ODD_preds$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
## Reference
## Prediction  No  Yes
##   No  6520 1054
##   Yes 2521  908
## Accuracy : 0.6751
## 95% CI : (0.6662, 0.6838)
## No Information Rate : 0.8217
## P-Value [Acc > NIR] : 1
## Kappa : 0.1423
## Mcnemar's Test P-Value : <2e-16
## Sensitivity : 0.46279
## Specificity : 0.72116
## Pos Pred Value : 0.26480
```
## Neg Pred Value : 0.86084
## Prevalence : 0.17832
## Detection Rate : 0.08252
## Detection Prevalence : 0.31164
## Balanced Accuracy : 0.59198
## 'Positive' Class : Yes

```
class_pred <- predict(rf_fit, test.set)
caret::confusionMatrix(class_pred$pred_class, test.set$fatal_od_case, positive = "Yes")
```

## Confusion Matrix and Statistics
## Reference
## Prediction  No  Yes
##        No  6520 1054
##        Yes 2521  908
##
## Accuracy : 0.6751
## 95% CI : (0.6662, 0.6838)
## No Information Rate : 0.8217
## P-Value [Acc > NIR] : 1
##
## Kappa : 0.1423
##
## McNemar's Test P-Value : <2e-16
##
## Sensitivity : 0.46279
## Specificity : 0.72116
## Pos Pred Value : 0.26480
## Neg Pred Value : 0.86084
## Prevalence : 0.17832
## Detection Rate : 0.08252
## Detection Prevalence : 0.31164
## Balanced Accuracy : 0.59198
##
## 'Positive' Class : Yes

### Oversample 1:1

```
ext oversample_rose_train_1 <- read_csv("U:/Andy's Thesis/MLoverdose/oversample_rose_train_1.csv", show_col_types = FALSE)
ext oversample_rose_train_1 <- oversample_rose_train_1 %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))

train.set <- oversample_rose_train_1
test.set <- datasettest
```

```
train.set = train.set %>% mutate_if(is.character, as.factor)
```
test.set = test.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 42230   50

dim(test.set)
## [1] 11003   50

nrow(train.set)
## [1] 42230

# XGBoost

xgb_spec <- boost_tree(
  trees = 300) %>%
  set_engine("xgboost") %>%
  set_mode("classification")

overdose_rec <- recipe(fatal_od_case ~., data = train.set)

xgbworkflow <- workflow() %>%
  add_model(xgb_spec) %>%
  add_recipe(overdose_rec)

xgb_fit <- xgbworkflow %>%
  fit(data = train.set)

class_pred <- predict(xgb_fit, test.set)
prob_pred <- predict(xgb_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("XGB_Class", "LR_DeathProb", "R_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)

caret::confusionMatrix(ODD_preds$XGB_Class, ODD_preds$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
## Reference
## Prediction  No  Yes
##   No 6454 1070
##   Yes 2587  892
##
## Accuracy : 0.6676
## 95% CI : (0.6587, 0.6764)
## No Information Rate : 0.8217
## P-Value [Acc > NIR] : 1
class_pred <- predict(xgb_fit, test.set)
caret::confusionMatrix(class_pred$response, test.set$fatal_od_case, positive = "Yes")

#Random Forest

rf_mod <-
  rand_forest(trees=500) %>%
set_engine("ranger") %>%
set_mode("classification")

rf_fit <-
rf_mod %>%
fit(fatal_od_case ~., data = train.set)

class_pred <- predict(rf_fit, test.set)
prob_pred <- predict(rf_fit, test.set, type = "prob")
predictions <- data.frame(class_pred, prob_pred) %>%
setNames(c("RF_Class", "RF_DeathProb", "RF_NotDeathProb"))

ODD_preds <- test.set %>%
bind_cols(predictions)
caret::confusionMatrix(ODD_preds$RF_Class, ODD_preds$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
##             Reference
## Prediction   No  Yes
##        No  6927 1195
##        Yes 2114  767
##
##                    Accuracy : 0.6993
##                       95% CI : (0.6906, 0.7078)
##                No Information Rate : 0.8217
##          P-Value [Acc > NIR] : 1
##
##                    Kappa : 0.1328
##
## Mcnemar's Test P-Value : <2e-16
##
##                    Sensitivity : 0.39093
##                    Specificity : 0.76618
##                Pos Pred Value : 0.26623
##                Neg Pred Value : 0.85287
##                    Prevalence : 0.17832
##            Detection Rate : 0.06971
##      Detection Prevalence : 0.26184
##            Balanced Accuracy : 0.57855
##
## 'Positive' Class : Yes
##

class_pred <- predict(rf_fit, test.set)
caret::confusionMatrix(class_pred$pred_class, test.set$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
##             Reference
## Prediction   No  Yes

test.set <- test.set %>%
  select(fatal_od_case, total_od_n, alzheimer_dementia_date_b, ami_date_b, angina_date_b, ckd_date_b, copd_date_b, depression_date_b, diabetes_date_b, heart_failure_date_b, hypertension_date_b, hosp_stroke_date_b, haemorr_stroke_date_b, isch_stroke_date_b, ihd_date_b, mood_anx_date_b, osteo_arthritis_date_b, osteo_porosis_date_b, ost, substance_related_disorders.x, Tobacco.use, Cocaine.use, earlyonsetdisorders, Mooddisorders, Multiplementalillness, Neurocognitivedisorders, Neuroticrelateddisorders, Opioiduse, Osteomyelitis, Otherpsychoactivedruguse, Personalitydisorders, Polysubstance, psychoticdisorders, Sepsis, Stimulantuse, Tissueinfection)

rf_mod <-
  rand_forest(trees=500) %>%
  set_engine("ranger") %>%
  set_mode("classification")

rf_fit <-
  rf_mod %>%
  fit(fatal_od_case ~., data = train.set)

class_pred <- predict(rf_fit, test.set)
prob_pred <- predict(rf_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("RF_Class", "RF_DeathProb", "RF_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)
caret::confusionMatrix(ODD_preds$RF_Class, ODD_preds$fatal_od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference
## Prediction  No  Yes
##    No     6466 1044
##   Yes     2575  918
##
##  Accuracy : 0.6711
##  95% CI : (0.6622, 0.6799)
##  No Information Rate : 0.8217
##  P-Value [Acc > NIR] : 1
##
##  Kappa : 0.1402
##
## McNemar's Test P-Value : <2e-16
##
##  Sensitivity : 0.46789
##  Specificity : 0.71519
##  Pos Pred Value : 0.26281
##  Neg Pred Value : 0.86099
Validating all the models

bothsample_rose_train_1 <- read_csv("U:/Andy's Thesis/MLoverdose/bothsample_rose_train_1.csv", show_col_types = FALSE)

bothsample_rose_train_1 <- bothsample_rose_train_1 %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))

datasettest <- read_csv("U:/Andy's Thesis/MLoverdose/test.set.csv", show_col_types = FALSE)
datasettest <- datasettest %>%
  mutate(fatal_od_case = recode(fatal_od_case, "0" = "No", "1" = "Yes"))

train.set <- bothsample_rose_train_1
test.set <- datasettest

train.set = train.set %>% mutate_if(is.character, as.factor)
test.set = test.set %>% mutate_if(is.character, as.factor)

#Xgboost

ten_fold <- vfold_cv(train.set, v=10)
xgb_spec <- boost_tree(
  trees = 300) %>%
  set_engine("xgboost") %>%
  set_mode("classification")

overdose_rec <-
  recipe(fatal_od_case ~., data = train.set)

xgbworkflow <- workflow() %>%
```r
add_model(xgb_spec) %>%
add_recipe(overdose_rec)

xgb_fit <- xgbworkflow %>%
  fit(data = train.set)

xgb_fit_rs <-
  xgbworkflow %>%
  fit_resamples(ten_fold)

collect_metrics(xgb_fit_rs)

## # A tibble: 2 x 5
## #  .metric .estimator  mean     n std_err
## 1 accuracy binary     0.754    10 0.00194
## 2 roc_auc  binary     0.839    10 0.00211

#Random Forest

ten_fold <- vfold_cv(train.set, v=10)

rf_spec <- rand_forest(
trees = 500) %>%
  set_engine("ranger") %>%
  set_mode("classification")

overdose_rec <-
  recipe(fatal_od_case ~., data = train.set)

rfworkflow <- workflow() %>%
  add_model(rf_spec) %>%
  add_recipe(overdose_rec)

rf_fit <- rfworkflow %>%
  fit(data = train.set)

rf_fit_rs <-
  rfworkflow %>%
  fit_resamples(ten_fold)

collect_metrics(rf_fit_rs)

## # A tibble: 2 x 5
## #  .metric .estimator  mean     n std_err
## 1 accuracy binary     0.799    10 0.00296
## 2 roc_auc  binary     0.883    10 0.00240

#Logistic Regression

ten_fold <- vfold_cv(train.set, v=10)
```
glm_spec <- logistic_reg() %>%
  set_engine("glm") %>%
  translate()

overdose_rec <-
  recipe(fatal_od_case ~ ., data = train.set)

glmworkflow <- workflow() %>%
  add_model(glm_spec) %>%
  add_recipe(overdose_rec)

glm_fit <- glmworkflow %>%
  fit(data = train.set)

glm_fit_rs <-
  glmworkflow %>%
  fit_resamples(ten_fold)

collect_metrics(glm_fit_rs)

## # A tibble: 2 x 5
##   .metric  .estimator  mean     n std_err
##   <chr>    <chr>      <dbl> <int>   <dbl>
## 1 accuracy binary     0.606    10 0.00246
## 2 roc_auc  binary     0.646    10 0.00165

#KNN

ten_fold <- vfold_cv(train.set, v=10)

knn_spec <- nearest_neighbor() %>%
  set_engine("kknn") %>%
  set_mode("classification") %>%
  translate

overdose_rec <-
  recipe(fatal_od_case ~ ., data = train.set)

knnworkflow <- workflow() %>%
  add_model(knn_spec) %>%
  add_recipe(overdose_rec)

knn_fit <- knnworkflow %>%
  fit(data = train.set)

knn_fit_rs <-
  knnworkflow %>%
  fit_resamples(ten_fold)

collect_metrics(knn_fit_rs)
## # A tibble: 2 x 5
## # .metric  .estimator  mean     n std_err
## <chr>    <chr>      <dbl> <int>   <dbl>
## 1 accuracy binary     0.766    10 0.00252
## 2 roc_auc  binary     0.757    10 0.00365

# Testing for the best performing model

### XGBoost
class_pred <- predict(xgb_fit, test.set)
prob_pred <- predict(xgb_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("XGB_Class", "LR_DeathProb", "R_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)

cmxgboost <- caret::confusionMatrix(ODD_preds$XGB_Class, ODD_preds$sal_od_case, positive = "Yes")

print(cmxgboost)

## Confusion Matrix and Statistics
## Reference
## Prediction  No  Yes
##       No  6117  972
##       Yes 2924  990
##
##          Accuracy : 0.6459
##          95% CI : (0.6369, 0.6549)
##     No Information Rate : 0.8217
##     P-Value [Acc > NIR] : 1
##
##          Kappa : 0.1304
##
## Mcnemar's Test P-Value : <2e-16
##
##          Sensitivity : 0.50459
##          Specificity : 0.67658
##          Pos Pred Value : 0.25294
##          Neg Pred Value : 0.86289
##          Prevalence : 0.17832
##          Detection Rate : 0.08998
##          Detection Prevalence : 0.35572
##          Balanced Accuracy : 0.59059
##
## 'Positive' Class : Yes
### RF

class_pred <- predict(rf_fit, test.set)
prob_pred <- predict(rf_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("rf_Class", "LR_DeathProb", "R_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)

cmrf <- caret::confusionMatrix(ODD_preds$rf_Class, ODD_preds$fatal_od_case, positive = "Yes")

print(cmrf)

## Confusion Matrix and Statistics
## Reference
## Prediction  No  Yes
##        No  6501 1062
##        Yes 2540  900
##
##                Accuracy : 0.6726
##                  95% CI : (0.6638, 0.6814)
##     No Information Rate : 0.8217
##     P-Value [Acc > NIR] : 1
##
##                Kappa : 0.1373
##
##     McNemar's Test P-Value : <2e-16
##
##    Sensitivity : 0.4587
##    Specificity : 0.7191
##    Pos Pred Value : 0.2616
##    Neg Pred Value : 0.8596
##    Prevalence : 0.1783
##    Detection Rate : 0.0818
##    Detection Prevalence : 0.3126
##    Balanced Accuracy : 0.5889
##
## 'Positive' Class : Yes
##
### GLM

class_pred <- predict(glm_fit, test.set)
prob_pred <- predict(glm_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("glm_Class", "LR_DeathProb", "R_NotDeathProb"))
ODD_preds <- test.set %>%
  bind_cols(predictions)

cm <- caret::confusionMatrix(ODD_preds$glm_Class, ODD_preds$fatal_od_case, positive = "Yes")

print(cm)

## Confusion Matrix and Statistics
##    Reference
## Prediction  No  Yes
##        No  5368  740
##        Yes 3673 1222
##
##                Accuracy : 0.5989
##                  95% CI : (0.5897, 0.6081)
##     No Information Rate : 0.8217
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1366
##
##  McNemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6228
##             Specificity : 0.5937
##          Pos Pred Value : 0.2496
##          Neg Pred Value : 0.8788
##              Prevalence : 0.1783
##          Detection Rate : 0.1111
##    Detection Prevalence : 0.4449
##       Balanced Accuracy : 0.6083
##
##     'Positive' Class : Yes
##
### KNN

class_pred <- predict(knn_fit, test.set)
prob_pred <- predict(knn_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("knn_Class", "LR_DeathProb", "R_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)

knncm <- caret::confusionMatrix(ODD_preds$knn_Class, ODD_preds$fatal_od_case, positive = "Yes")

print(knncm)
## Confusion Matrix and Statistics
### Reference
### Prediction  No  Yes
### No  7257  1493
### Yes  1784  469
###
### Accuracy : 0.7022
### 95% CI : (0.6935, 0.7107)
### No Information Rate : 0.8217
### P-Value [Acc > NIR] : 1
###
### Kappa : 0.0394
###
### McNemar's Test P-Value : 4.064e-07
###
### Sensitivity : 0.23904
### Specificity : 0.80268
### Pos Pred Value : 0.20817
### Neg Pred Value : 0.82937
### Prevalence : 0.17832
### Detection Rate : 0.04262
### Detection Prevalence : 0.20476
### Balanced Accuracy : 0.52086
###
### 'Positive' Class : Yes
###

Inference code

### Shap XGBOOST

```r
library(data.table)
library(tidyverse)
```

```r
## v dplyr 1.1.2 v readr 2.1.4
## v forcats 1.0.0 v stringr 1.5.0
## v ggplot2 3.4.2 v tibble 3.2.1
## v lubridate 1.9.2 v tidyr 1.3.0
## v purrr 1.0.1
```

```r
## x dplyr::between() masks data.table::between()
## x dplyr::filter() masks stats::filter()
## x dplyr::first() masks data.table::first()
## x lubridate::hour() masks data.table::hour()
## x lubridate::isoweek() masks data.table::isoweek()
## x dplyr::lag() masks stats::lag()
## x dplyr::last() masks data.table::last()
## x lubridate::mday() masks data.table::mday()
```
## x lubridate::minute() masks data.table::minute()
## x lubridate::month() masks data.table::month()
## x lubridate::quarter() masks data.table::quarter()
## x lubridate::second() masks data.table::second()
## x purrr::transpose() masks data.table::transpose()
## x lubridate::wday() masks data.table::wday()
## x lubridate::week() masks data.table::week()
## x lubridate::yday() masks data.table::yday()
## x lubridate::year() masks data.table::year()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(caret)

## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##       lift

library(xgboost)

##
## Attaching package: 'xgboost'
##
## The following object is masked from 'package:dplyr':
##
##       slice

library(SHAPforxgboost)
library(here)

## here() starts at R:/atai-22-075/Andy's Thesis/MLoverdose

library(ggplot2)
library(Boruta)
library(randomForest)

## randomForest 4.6-14
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##       combine
##
## The following object is masked from 'package:ggplot2':
##
##       margin

library(naniar)
datasetmlmodel2 <- read_csv("U:/Andy's Thesis/MLoverdose/datasetmlmodel2.csv", show_col_types = FALSE)

table(datasetmlmodel2$fatal_od_case)

##
##     0     1
## 30253  6426

gg_miss_var(datasetmlmodel2)

## Warning: The `guide` argument in `scale_*()` cannot be `FALSE`. This was deprecated in
## ggplot2 3.3.4.
## i Please use "none" instead.
## i The deprecated feature was likely used in the naniar package.
## Please report the issue at <https://github.com/njtierney/naniar/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::.last_lifecycle_warnings()` to see where this warning was
## generated.
datasetmlmodel3 <- datasetmlmodel2 %>%
  select(-c(first_od_date, last_od_date))

Logit Regression
glmmodel <- glm(fatal_od_case ~ ., data = datasetmlmodel3, family = binomial(link = "logit"))

summary(glmmodel)

## Call:
## glm(formula = fatal_od_case ~ ., family = binomial(link = "logit"),
##     data = datasetmlmodel3)
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.5413 -0.6824  0.5292  0.3658  2.7721
##
## Coefficients:

|                      | Estimate | Std. Error | z value | Pr(>|z|)  |
|----------------------|----------|------------|---------|----------|
| (Intercept)          | -0.952975| 0.079829   | -11.938 | < 2e-16  *** |
| total_od_n           | 0.027188 | 0.008003   | 3.397   | 0.000681 *** |
| alzheimer_dementia_b | -0.377881| 0.114914   | -3.288  | 0.001008 ** |
| ami_date_b           | 0.082794 | 0.102636   | 0.807   | 0.419856 |
| angina_date_b        | -0.231245| 0.090551   | -2.549  | 0.010749  |
| asthma_date_b        | -0.079273| 0.032731   | -2.434  | 0.014900  |
| ckd_date_b           | 0.118923 | 0.053099   | 2.238   | 0.024765  |
| copd_date_b          | 0.260162 | 0.051363   | 5.078   | < 2e-16  *** |
| depression_date_b    | 0.416079 | 0.056403   | 7.377   | < 2e-16  *** |
| diabetes_date_b      | 0.004663 | 0.049939   | 0.093   | 0.925606 |
| epilepsy_date_b      | 0.002314 | 0.073062   | 0.032   | 0.974730 |
| heart_failure_date_b | 0.379520 | 0.068265   | 5.560   | < 2e-16  *** |
| hypertension_date_b  | 0.251090 | 0.048101   | 5.218   | < 2e-16  *** |
| hosp_stroke_date_b   | 0.719826 | 0.492474   | 1.462   | 0.143836 |
| haemorr_stroke_date_b| 0.195340 | 0.470162   | 0.415   | 0.677795 |
| isch_stroke_date_b   | -0.511212| 0.481887   | -1.061  | 0.288756 |
| hosp_tia_date_b      | -0.281877| 0.218268   | -1.291  | 0.196555 |
| ihd_date_b           | 0.015880 | 0.069179   | -0.230  | 0.818448 |
| mood_anx_date_b      | 0.082597 | 0.059264   | 1.394   | 0.163407 |
| ms_date_b            | -0.167895| 0.252270   | -0.666  | 0.505708 |
| osteo_arthritis_date_b| 0.077570| 0.044958   | 1.725   | 0.084459 |
| osteo_porosis_date_b | -0.271311| 0.090551   | -2.996  | 0.002734 ** |
| parkinsonism_date_b  | 0.414190 | 0.247560   | 1.673   | 0.094309 |
| rheumatoid_arthritis_date_b| -0.133540| 0.107213| -1.246  | 0.212926 |
| oat                  | -0.552500| 0.038274   | -14.435 | < 2e-16  *** |
| substancerelateddisorders| -0.201061| 0.038552| -5.215  | 1.84e-07 *** |
| Tobaccouse           | -0.026204| 0.068409   | -0.426  | 0.669396 |
| Alcoholuse           | -0.085234| 0.080235   | -1.062  | 0.288100 |
| Behaviouralpsychologicaldisturbances| -0.282280| 0.151576| -1.862  | 0.062561 . |
| Cannabinoiduse       | -0.012893| 0.052625   | -0.245  | 0.806466 |
| Cocaineuse           | 0.457615 | 0.042044   | 10.884  | < 2e-16  *** |
| Developmentdisorders | -0.040587| 0.100026   | -0.406  | 0.684917 |
| earlyonsetdisorders  | -0.304941| 0.042100   | -7.243  | 4.38e-13 *** |
| Endocarditis         | -0.020175| 0.085544   | -0.236  | 0.813552 |
| Hallucinogensuse     | -0.413666| 0.199469   | -2.077  | 0.037770 * |
| Intellectualdisability| -0.438217| 0.210483| -2.082  | 0.037346 * |
| Mooddisorders        | -0.105439| 0.037959   | -2.778  | 0.005474 ** |
| Multiplementalillness| -0.109086| 0.032982   | -3.307  | 0.000941 *** |
| Neurocogmitivedisorders| -0.117844| 0.046509| -2.534  | 0.011284 * |
| Neuroticrelateddisorders| -0.325998| 0.033587| -9.706  | < 2e-16  *** |
| Opioiduse            | -0.313626| 0.037175   | -8.437  | < 2e-16  *** |
| Osteomyelitis        | -0.100375| 0.075871   | -1.323  | 0.185847 |
| Otherpsychoactivedruguse| -0.416606| 0.050841| -8.194  | 2.52e-16 *** |
| Personalitydisorders | 0.065353 | 0.044640   | 1.464   | 0.143197 |
| Polysubstance        | -0.102430| 0.038804   | -2.604  | 0.008299 ** |
| Psychoticdisorders   | 0.027167 | 0.044628   | 0.609   | 0.542704 |
| Sedativeandhypnoticuse| 0.187529| 0.081701| 2.295   | 0.021715 * |
| Sepsis               | -0.067791| 0.045980   | -1.474  | 0.140387 |
| Stimulantuse         | -0.209336| 0.050368   | -4.156  | 3.24e-05 *** |
## Tissueinfection

-0.177368   0.038262   4.636 3.56e-06 ***

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## (Dispersion parameter for binomial family taken to be 1)

## Null deviance: 34040  on 36678  degrees of freedom
## Residual deviance: 32159  on 36629  degrees of freedom
## AIC: 32259

## Number of Fisher Scoring iterations: 5

plot(glmmodel)
# oddsratio <- exp(cbind(Odds_Ratio = coef(glmmodel), confint(glmmodel)))
#
# oddsratio
```{r}
# oddsratio1 <- as.data.frame(oddsratio)
# write.csv(oddsratio, "oddsratio.csv")

oddsratio <- read_csv("U:/Andy's Thesis/MLoverdose/oddsratio.csv", show_col_types = FALSE)

oddsratio

## # A tibble: 50 x 4
## # Variables Cadle_Ratio SDE2.5 SDE975
## # <chr> <dbl> <dbl> <dbl>
## 1 Acute_Myocardial_Infraction 1.09 0.887 1.33
## 2 Alcohol_use 0.918 0.786 1.08
## 3 Alzheimer_Dementia 0.685 0.545 0.855
## 4 Angina 0.794 0.648 0.969
## 5 Asthma 0.924 0.864 0.987
## 6 Behavioural_psychological_disturbances 0.754 0.554 1.01
## 7 Cannabinoid_use 0.987 0.890 1.09
## 8 Chronic_Obstructive_Pulmonary_Disease 1.30 1.17 1.43
## 9 Chronic_Kidney_Disease 1.13 1.01 1.25
## 10 Cocaine_use 1.58 1.46 1.72
## # i 40 more rows

oddsratio %>%
  print(n=50)

## # A tibble: 50 x 4
## # Variables Cadle_Ratio SDE2.5 SDE975
## # <chr> <dbl> <dbl> <dbl>
## 1 Acute_Myocardial_Infraction 1.09 0.887 1.33
## 2 Alcohol_use 0.918 0.786 1.08
## 3 Alzheimer_Dementia 0.685 0.545 0.855
## 4 Angina 0.794 0.648 0.969
## 5 Asthma 0.924 0.864 0.987
## 6 Behavioural_psychological_disturbances 0.754 0.554 1.01
## 7 Cannabinoid_use 0.987 0.890 1.09
## 8 Chronic_Obstructive_Pulmonary_Disease 1.30 1.17 1.43
## 9 Chronic_Kidney_Disease 1.13 1.01 1.25
## 10 Cocaine_use 1.58 1.46 1.72
## 11 Depression 1.52 1.36 1.69
## 12 Development_disorders 0.964 0.746 1.16
## 13 Diabetes 1.00 0.911 1.11
## 14 Earlyonset_disorders 0.737 0.678 0.800
## 15 Endocarditis 0.980 0.827 1.16
## 16 Epilepsy 1.00 0.867 1.15
## 17 Haemorrhagic_Stroke 1.22 0.454 2.95
## 18 Hallucinogens_use 0.661 0.438 0.960
## 19 Heart_Failure 1.46 1.28 1.67
## 20 Hospital_Stroke 2.05 0.808 5.71
```
<table>
<thead>
<tr>
<th>Variable</th>
<th>Lower CI</th>
<th>Estimate</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>#21 Hospital_TIA</td>
<td>0.754</td>
<td>1.14</td>
<td>0.483</td>
</tr>
<tr>
<td>#22 Hypertension</td>
<td>1.29</td>
<td>1.41</td>
<td>1.17</td>
</tr>
<tr>
<td>#23 Intellectual_disability</td>
<td>0.645</td>
<td>0.957</td>
<td>0.418</td>
</tr>
<tr>
<td>#24 Intercept</td>
<td>0.386</td>
<td>0.450</td>
<td>0.329</td>
</tr>
<tr>
<td>#25 Ischemic_stroke</td>
<td>0.984</td>
<td>1.13</td>
<td>0.859</td>
</tr>
<tr>
<td>#26 Ischmeic_Heart_Disease</td>
<td>0.600</td>
<td>1.49</td>
<td>0.220</td>
</tr>
<tr>
<td>#27 Mood_Anxiety</td>
<td>0.921</td>
<td>1.03</td>
<td>0.819</td>
</tr>
<tr>
<td>#28 Mood_disorders</td>
<td>0.900</td>
<td>0.969</td>
<td>0.835</td>
</tr>
<tr>
<td>#29 Multiple_mental_illness</td>
<td>0.897</td>
<td>0.957</td>
<td>0.841</td>
</tr>
<tr>
<td>#30 Multiple_Sclerosis</td>
<td>0.845</td>
<td>1.13</td>
<td>0.839</td>
</tr>
<tr>
<td>#31 Neurocognitive_disorders</td>
<td>0.889</td>
<td>0.973</td>
<td>0.811</td>
</tr>
<tr>
<td>#32 Neuroticrelated_disorders</td>
<td>0.722</td>
<td>0.771</td>
<td>0.676</td>
</tr>
<tr>
<td>#33 OAT</td>
<td>0.973</td>
<td>1.01</td>
<td>0.973</td>
</tr>
<tr>
<td>#34 Opioid_use</td>
<td>0.644</td>
<td>0.957</td>
<td>0.518</td>
</tr>
<tr>
<td>#35 Osteomyelitis</td>
<td>0.904</td>
<td>1.05</td>
<td>0.778</td>
</tr>
<tr>
<td>#36 Osteoporosis</td>
<td>0.762</td>
<td>0.909</td>
<td>0.637</td>
</tr>
<tr>
<td>#37 Other_psychoactive_drug_use</td>
<td>0.659</td>
<td>0.728</td>
<td>0.596</td>
</tr>
<tr>
<td>#38 Parkinsonism</td>
<td>1.51</td>
<td>2.42</td>
<td>0.915</td>
</tr>
<tr>
<td>#39 Personality_disorders</td>
<td>1.07</td>
<td>1.41</td>
<td>0.978</td>
</tr>
<tr>
<td>#40 Polysubstance</td>
<td>1.03</td>
<td>1.12</td>
<td>0.941</td>
</tr>
<tr>
<td>#41 Psychoticdisorders</td>
<td>0.762</td>
<td>0.990</td>
<td>0.657</td>
</tr>
<tr>
<td>#42 Rheumatoid_Arthritis</td>
<td>1.21</td>
<td>1.41</td>
<td>1.03</td>
</tr>
<tr>
<td>#43 Sedative_and_hypnoticuse</td>
<td>0.973</td>
<td>1.10</td>
<td>0.855</td>
</tr>
<tr>
<td>#44 Sepsis</td>
<td>0.973</td>
<td>1.10</td>
<td>0.855</td>
</tr>
<tr>
<td>#45 Sedative and hypnoticuse</td>
<td>0.973</td>
<td>1.10</td>
<td>0.855</td>
</tr>
<tr>
<td>#46 Stimulant_use</td>
<td>0.973</td>
<td>1.10</td>
<td>0.855</td>
</tr>
<tr>
<td>#47 Substanrelated Disorders</td>
<td>0.973</td>
<td>1.10</td>
<td>0.855</td>
</tr>
<tr>
<td>#48 Tissue_infection</td>
<td>0.973</td>
<td>1.10</td>
<td>0.855</td>
</tr>
<tr>
<td>#49 Tobacco_use</td>
<td>0.973</td>
<td>1.10</td>
<td>0.855</td>
</tr>
<tr>
<td>#50 Total Overdose</td>
<td>0.973</td>
<td>1.10</td>
<td>0.855</td>
</tr>
</tbody>
</table>

```
ggplot(oddsratio, aes(x = Odds_Ratio, y = reorder(Variables) + Odds_Ratio)) +
  geom_point(stat = "identity", shape = 15) +
  geom_line(yintercept = 1, linetype = "dashed", color = "grey") +
  geom_errorbar(aes(xmin = SDE2.5, xmax = SDE975, width = 0.4))
```
#XGboost’s Shap Values

datasetmlmodel31 <- datasetmlmodel3[,-1]
dataX <- as.matrix(datasetmlmodel31)

param_list <- list(objective = "reg:squarederror",
                    eta = 0.02,
                    max_depth = 10,
                    gamma = 0.01,
                    subsample = 0.95)

mod <- xgboost::xgboost(data = dataX,
                        label = as.matrix(datasetmlmodel3$fatal_od_case),
                        params = param_list,
                        nrounds = 10,
                        verbose = FALSE,
                        nthread = parallel::detectCores() - 2,
                        early_stopping_rounds = 8)

shap_values <- shap.values(xgb_model = mod, X_train = dataX)

shap_long <- shap.prep(shap_contrib = shap_values$shap_score, X_train = dataX)

shap.plot.summary(shap_long)
shap_int <- shap.prep.interaction(xgb_mod = mod, X_train = dataX)

shap_values$mean_shap_score

##                         oat             Neuroticrelateddisorders
##                         6.581000e-03                         4.550315e-03
##                            Opioiduse            substancerelateddisorders
##                         4.332968e-03                         2.598550e-03
##             Otherpsychoactivedruguse                           total_od_n
##                         2.367992e-03                         2.309789e-03
##                        Polysubstance                    depression_date_b
##                         1.885902e-03                         1.796028e-03
##                           Cocaineuse                  earlyonsetdisorders
##                         1.688843e-03                         1.484034e-03
##                  hypertension_date_b                      Tissueinfection
##                         1.209806e-03                         1.144584e-03
##                  heart_failure_date_b                      Stimulantuse
##                         1.072378e-03                         6.591153e-04
##                         copd_date_b                           ckd_date_b
##                         6.429045e-04                         5.887527e-04
##                          copd_date_b                           ckd_date_b
##                         5.170550e-04                         4.595692e-04
##                          copd_date_b                           ckd_date_b
##                         4.079154e-04                         3.597776e-04
##                          copd_date_b                           ckd_date_b
##                         6.429045e-04                         5.887527e-04
##                          copd_date_b                           ckd_date_b
##                         3.350496e-04                         3.044582e-04
##                         2.778263e-04                         2.564722e-04
##                         2.107182e-04                         1.938154e-04
##                         1.901489e-04                         1.894970e-04
##                         1.854932e-04                         1.831324e-04
##                         1.789100e-04                         1.678846e-04
##                         1.615564e-04                         1.143478e-04
##                          1.19513e-04                         9.890473e-05
##                          Endocarditis          rheumatoid_arthritis_date_b
##                         8.745864e-05                         8.352768e-05
##                         Osteomyelitis                           Tobaccouse
##                         8.023228e-05                         7.759541e-05
##                         isch_stroke_date_b                 Developmentdisorders
##                         7.486815e-05                         5.592261e-05
##                         hosp_tia_date_b                  parkinsonism_date_b
##                         3.840452e-05                         3.747708e-05
##                         Hallucinogensuse Behaviouralpsychologicaldisturbances
##                         3.262430e-05                         2.973329e-05
##                         Alcoholuse                       ms_date_b
##                         1.871776e-05                         1.629704e-05
##                         Intellectualdisability
##                         1.359569e-05

g1 <- shap.plot.dependence(data_long = shap_long, 
data_int = shap_int, 
x= "total_od_n", 
y = "Polysubstance", 
color_feature = "Polysubstance")
g2 <- shap.plot.dependence(data_long = shap_long, 
data_int = shap_int, 
x= "total_od_n", 
y = "Mooddisorders", 
color_feature = "Mooddisorders")
g3 <- shap.plot.dependence(data_long = shap_long,
data_int = shap_int,
x = "total_od_n",
y = "Multiplementalillness",
color_feature = "Multiplementalillness")
gridExtra::grid.arrange(g1, g2, g3, ncol=3)

## `geom_smooth()` using formula = y ~ x

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 0.64
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 1.36
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 1.2538e-026
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1
## `geom_smooth()` using formula = y ~ x

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 0.64
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 1.36
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 1.2538e-026
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1
## `geom_smooth()` using formula = y ~ x

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 0.64
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 1.36
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 1.2538e-026
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1
ggplot2

```r
## `geom_smooth()` using formula = 'y ~ x'
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 0.64
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 1.36
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 1.2538e-026
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1
```
g2

```r
## `geom_smooth()` using formula = 'y ~ x'
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 0.64
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 1.36
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 1.2538e-026
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1
```
```
g3

# `geom_smooth()` using formula = 'y ~ x'

# Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
# parametric, : pseudoinverse used at 0.64

# Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
# parametric, : neighborhood radius 1.36

# Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
# parametric, : reciprocal condition number 1.2538e-026

# Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
# parametric, : There are other near singularities as well. 1
```
### Variable importance from Boruta

```r
table(datasetmlmodel3$fatal_od_case)
```

```r
# Borutamodel_v1 <- Boruta(fatal_od_case ~., data = datasetmlmodel3, doTrace = 2)

# saveRDS(Borutamodel, file = "borutamodel_V1.rds")

plot1 <- plot(Borutamodel_v1, xlab = "", xaxt = "n")
```


**getSelectedAttributes**(*Borutamodel_v1, withTentative = F*)

```r
## [1] "total_od_n"                  "alzheimer_dementia_date_b"
## [3] "ami_date_b"                   "angina_date_b"
## [5] "asthma_date_b"                "ckd_date_b"
## [7] "copd_date_b"                  "depression_date_b"
## [9] "diabetes_date_b"              "heart_failure_date_b"
## [11] "hypertension_date_b"          "hosp_stroke_date_b"
## [13] "haemorr_stroke_date_b"        "isch_stroke_date_b"
## [15] "hosp_tia_date_b"              "ihd_date_b"
## [17] "mood_anx_date_b"             "osteo_arthritis_date_b"
## [19] "osteo_porosis_date_b"         "rheumatoid_arthritis_date_b"
## [21] "oat"                          "substancerelateddisorders"
## [23] "Tobaccouse"                   "Cocaineuse"
## [25] "earlyonsetdisorders"          "Endocarditis"
## [27] "Mooddisorders"                "Multiplementalillness"
## [29] "Neurocognitivedisorders"      "Neuroticrelateddisorders"
## [31] "Opioiduse"                    "Osteomyelitis"
## [33] "Otherpsychoactivedruguse"     "Personalitydisorders"
## [35] "Polysubstance"                "psychoticdisorders"
## [37] "Sedativeandhypnoticuse"       "Sepsis"
## [39] "Stimulantuse"                 "Tissueinfection"
```

### Variable importance from caret

```r
rPartMod <- train(fatal_od_case ~ ., data = datasetmlmodel3, method = "rpart")
```
## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to do
## classification? If so, use a 2 level factor as your outcome column.

## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo, :
## There were missing values in resampled performance measures.

rpartImp <- varImp(rPartMod)

print(rpartImp)

## rpart variable importance

## only 20 most important variables shown (out of 49)

## Overall
## Opioiduse                             100.00
## substancerelateddisorders              93.05
## Neuroticrelateddisorders               92.42
## Polysubstance                          87.67
## oat                                    76.45
## Otherpsychoactivedruguse               28.56
## Multiplementalillness                  0.00
## hypertension_date_b                    0.00
## ami_date_b                             0.00
## Tobaccouse                             0.00
## ihd_date_b                             0.00
## Intellectualdisability                 0.00
## Developmentdisorders                   0.00

roc_imp <- filterVarImp(x = datasetmlmodel3[,2:50], y = datasetmlmodel3$fatal_od_case)

roc_imp <- data.frame(cbind(variable = rownames(roc_imp), score = roc_imp[,1]))

roc_imp$score <- as.double(roc_imp$score)

library(dplyr)
rocscore <- roc_imp %>% arrange(desc(score))
print(rocscore, nrow(Inf))
<table>
<thead>
<tr>
<th></th>
<th>Category</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Neuroticrelateddisorders</td>
<td>19.76087956</td>
</tr>
<tr>
<td>2</td>
<td>Otherpsychoactivedruguse</td>
<td>19.04287068</td>
</tr>
<tr>
<td>3</td>
<td>earlyonsetdisorders</td>
<td>14.68934749</td>
</tr>
<tr>
<td>4</td>
<td>Stimulantuse</td>
<td>14.65597891</td>
</tr>
<tr>
<td>5</td>
<td>Mooddisorders</td>
<td>12.68422444</td>
</tr>
<tr>
<td>6</td>
<td>Multiplementalillness</td>
<td>12.26715826</td>
</tr>
<tr>
<td>7</td>
<td>hypertension_date_b</td>
<td>10.52028615</td>
</tr>
<tr>
<td>8</td>
<td>haemorr_stroke_date_b</td>
<td>9.15459231</td>
</tr>
<tr>
<td>9</td>
<td>heart_failure_date_b</td>
<td>9.10571467</td>
</tr>
<tr>
<td>10</td>
<td>hosp_stroke_date_b</td>
<td>8.86513144</td>
</tr>
<tr>
<td>11</td>
<td>Personalitydisorders</td>
<td>8.62889544</td>
</tr>
<tr>
<td>12</td>
<td>total_od_n</td>
<td>8.50911235</td>
</tr>
<tr>
<td>13</td>
<td>psychoticdisorders</td>
<td>7.91715709</td>
</tr>
<tr>
<td>14</td>
<td>copd_date_b</td>
<td>7.44417782</td>
</tr>
<tr>
<td>15</td>
<td>Alcoholuse</td>
<td>6.79141891</td>
</tr>
<tr>
<td>16</td>
<td>Cannabinoiuse</td>
<td>6.73367799</td>
</tr>
<tr>
<td>17</td>
<td>Sepsis</td>
<td>6.58966398</td>
</tr>
<tr>
<td>18</td>
<td>ihd_date_b</td>
<td>5.90571156</td>
</tr>
<tr>
<td>19</td>
<td>osteo_arthritis_date_b</td>
<td>5.52811728</td>
</tr>
<tr>
<td>20</td>
<td>diabetes_date_b</td>
<td>5.32382678</td>
</tr>
<tr>
<td>21</td>
<td>ami_date_b</td>
<td>5.11989841</td>
</tr>
<tr>
<td>22</td>
<td>isch_stroke_date_b</td>
<td>4.77846504</td>
</tr>
<tr>
<td>23</td>
<td>Neurocognitivedisorders</td>
<td>4.71830598</td>
</tr>
<tr>
<td>24</td>
<td>ckd_date_b</td>
<td>4.66400256</td>
</tr>
<tr>
<td>25</td>
<td>Developmentdisorders</td>
<td>4.44972851</td>
</tr>
<tr>
<td>26</td>
<td>Hallucinogensuse</td>
<td>4.07639176</td>
</tr>
<tr>
<td>27</td>
<td>Osteomyelitis</td>
<td>4.03697054</td>
</tr>
<tr>
<td>28</td>
<td>Tissueinfection</td>
<td>4.02613999</td>
</tr>
<tr>
<td>29</td>
<td>Intellectualdisability</td>
<td>3.87822689</td>
</tr>
<tr>
<td>30</td>
<td>Behaviouralpsychologicaldisturbances</td>
<td>3.81069677</td>
</tr>
<tr>
<td>31</td>
<td>Tobaccouse</td>
<td>3.63407191</td>
</tr>
<tr>
<td>32</td>
<td>mood_anx_date_b</td>
<td>3.21643091</td>
</tr>
<tr>
<td>33</td>
<td>angina_date_b</td>
<td>2.97314226</td>
</tr>
<tr>
<td>34</td>
<td>Sedativeandhypnoticuse</td>
<td>2.93817285</td>
</tr>
<tr>
<td>35</td>
<td>Endocarditis</td>
<td>2.32552949</td>
</tr>
<tr>
<td>36</td>
<td>asthma_date_b</td>
<td>2.20658461</td>
</tr>
<tr>
<td>37</td>
<td>parkinsonism_date_b</td>
<td>1.87384084</td>
</tr>
<tr>
<td>38</td>
<td>Cocaineuse</td>
<td>0.94083613</td>
</tr>
<tr>
<td>39</td>
<td>depression_date_b</td>
<td>0.77332786</td>
</tr>
<tr>
<td>40</td>
<td>alzheimer_dementia_date_b</td>
<td>0.57767102</td>
</tr>
<tr>
<td>41</td>
<td>osteo_porosis_date_b</td>
<td>0.51930807</td>
</tr>
<tr>
<td>42</td>
<td>hosp_tia_date_b</td>
<td>0.25468715</td>
</tr>
<tr>
<td>43</td>
<td>ms_date_b</td>
<td>0.20489051</td>
</tr>
<tr>
<td>44</td>
<td>rheumatoid_arthritis_date_b</td>
<td>0.08151329</td>
</tr>
<tr>
<td>45</td>
<td>epilepsy_date_b</td>
<td>0.07975992</td>
</tr>
</tbody>
</table>

```r
library(randomForest)
rf = randomForest(x = datasetmlmodel3[,2:50], y = datasetmlmodel3$fatal_od_case)
```
## Warning in randomForest.default(x = datasetmlmodel3[, 2:50], y =
## datasetmlmodel3$fatal_od_case): The response has five or fewer unique values.
## Are you sure you want to do regression?

```r
var_imp_rf <- varImp(rf, scale = FALSE)
```

```r
var_imp_rf1 <- var_imp_rf %>% arrange(desc(Overall))
view(var_imp_rf1, nrow(Inf))
```

## RFE Boosting Bagging to do feature selection

```r
filterCtrl <- rfeControl(functions = treebagFuncs, method = "cv", number = 10)
```

```r
boostingrfe <- rfe(x = datasetmlmodel3[,2:50], y = datasetmlmodel3$fatal_od_case, sizes = c(1:49), rfeControl = filterCtrl)
```

```r
boostingrfe
```

### Recursive feature selection

### Outer resampling method: Cross-Validated (10 fold)

### Resampling performance over subset size:

### The top 5 variables (out of 49):

```
oat, Opioiduse, substancerelateddisorders, Polysubstance, Neuroticrelateddisorders
```

```r
plot(boostingrfe, type = c("g", "o"))
```
predictors(boostergrfe)

## [1] "oat"
## [2] "Opioiduse"
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[3]</td>
<td>&quot;substancerelateddisorders&quot;</td>
</tr>
<tr>
<td>[4]</td>
<td>&quot;Polysubstance&quot;</td>
</tr>
<tr>
<td>[5]</td>
<td>&quot;Neuroticrelateddisorders&quot;</td>
</tr>
<tr>
<td>[6]</td>
<td>&quot;Otherpsychoactivedruguse&quot;</td>
</tr>
<tr>
<td>[7]</td>
<td>&quot;earlyonsetdisorders&quot;</td>
</tr>
<tr>
<td>[8]</td>
<td>&quot;Alcoholuse&quot;</td>
</tr>
<tr>
<td>[9]</td>
<td>&quot;alzheimer_dementia_date_b&quot;</td>
</tr>
<tr>
<td>[10]</td>
<td>&quot;ami_date_b&quot;</td>
</tr>
<tr>
<td>[11]</td>
<td>&quot;angina_date_b&quot;</td>
</tr>
<tr>
<td>[12]</td>
<td>&quot;asthma_date_b&quot;</td>
</tr>
<tr>
<td>[13]</td>
<td>&quot;Behaviouralpsychologicaldisturbances&quot;</td>
</tr>
<tr>
<td>[14]</td>
<td>&quot;Cannabinoiduse&quot;</td>
</tr>
<tr>
<td>[15]</td>
<td>&quot;ckd_date_b&quot;</td>
</tr>
<tr>
<td>[16]</td>
<td>&quot;Cocaineuse&quot;</td>
</tr>
<tr>
<td>[17]</td>
<td>&quot;copd_date_b&quot;</td>
</tr>
<tr>
<td>[18]</td>
<td>&quot;depression_date_b&quot;</td>
</tr>
<tr>
<td>[19]</td>
<td>&quot;Developmentdisorders&quot;</td>
</tr>
<tr>
<td>[20]</td>
<td>&quot;diabetes_date_b&quot;</td>
</tr>
<tr>
<td>[21]</td>
<td>&quot;Endocarditis&quot;</td>
</tr>
<tr>
<td>[22]</td>
<td>&quot;epilepsy_date_b&quot;</td>
</tr>
<tr>
<td>[23]</td>
<td>&quot;haemorr_stroke_date_b&quot;</td>
</tr>
<tr>
<td>[24]</td>
<td>&quot;Hallucinogensuse&quot;</td>
</tr>
<tr>
<td>[25]</td>
<td>&quot;heart_failure_date_b&quot;</td>
</tr>
<tr>
<td>[26]</td>
<td>&quot;hosp_stroke_date_b&quot;</td>
</tr>
<tr>
<td>[27]</td>
<td>&quot;hosp_tia_date_b&quot;</td>
</tr>
<tr>
<td>[28]</td>
<td>&quot;hypertension_date_b&quot;</td>
</tr>
<tr>
<td>[29]</td>
<td>&quot;ihd_date_b&quot;</td>
</tr>
<tr>
<td>[30]</td>
<td>&quot;Intellectualdisability&quot;</td>
</tr>
<tr>
<td>[31]</td>
<td>&quot;isch_stroke_date_b&quot;</td>
</tr>
<tr>
<td>[32]</td>
<td>&quot;mood_anx_date_b&quot;</td>
</tr>
<tr>
<td>[33]</td>
<td>&quot;Mooddisorders&quot;</td>
</tr>
<tr>
<td>[34]</td>
<td>&quot;ms_date_b&quot;</td>
</tr>
<tr>
<td>[35]</td>
<td>&quot;Multiplementalillness&quot;</td>
</tr>
<tr>
<td>[36]</td>
<td>&quot;Neurocognitivedisorders&quot;</td>
</tr>
<tr>
<td>[37]</td>
<td>&quot;osteo_arthritis_date_b&quot;</td>
</tr>
<tr>
<td>[38]</td>
<td>&quot;osteo_porosis_date_b&quot;</td>
</tr>
<tr>
<td>[39]</td>
<td>&quot;Osteomyelitis&quot;</td>
</tr>
<tr>
<td>[40]</td>
<td>&quot;parkinsonism_date_b&quot;</td>
</tr>
<tr>
<td>[41]</td>
<td>&quot;Personalitydisorders&quot;</td>
</tr>
<tr>
<td>[42]</td>
<td>&quot;psychoticdisorders&quot;</td>
</tr>
<tr>
<td>[43]</td>
<td>&quot;rheumatoid_arthritis_date_b&quot;</td>
</tr>
<tr>
<td>[44]</td>
<td>&quot;Sedativeandhypnoticuse&quot;</td>
</tr>
<tr>
<td>[45]</td>
<td>&quot;Sepsis&quot;</td>
</tr>
<tr>
<td>[46]</td>
<td>&quot;Stimulantuse&quot;</td>
</tr>
<tr>
<td>[47]</td>
<td>&quot;Tissueinfection&quot;</td>
</tr>
<tr>
<td>[48]</td>
<td>&quot;Tobaccouse&quot;</td>
</tr>
<tr>
<td>[49]</td>
<td>&quot;total_od_n&quot;</td>
</tr>
</tbody>
</table>
rfevarimp <- data.frame(varImp(boostingrfe))
rfevarimp

## Overall
## oat  0.0170208809
## Opioiduse  0.0153418276
## substance-related disorders  0.0137619873
## Polysubstance  0.0129361766
## Neurotic-related disorders  0.0087437654
## Other psychoactive drug use  0.0030738596
## early onset disorders  0.0000458704
## Alcohol use  0.0000000000
## alzheimer_dementia_date_b  0.0000000000
## ami_date_b  0.0000000000
## angina_date_b  0.0000000000
## asthma_date_b  0.0000000000
## Behavioural psychological disturbances  0.0000000000
## Cannabinoid use  0.0000000000
## ckd_date_b  0.0000000000
## Cocaine use  0.0000000000
## copd_date_b  0.0000000000
## depression_date_b  0.0000000000
## Development disorders  0.0000000000
## diabetes_date_b  0.0000000000
## Endocarditis  0.0000000000
## epilepsy_date_b  0.0000000000
## haemorrhage_stroke_date_b  0.0000000000
## Hallucinogens use  0.0000000000
## heart failure_date_b  0.0000000000
## hosp_stroke_date_b  0.0000000000
## hosp_tia_date_b  0.0000000000
## hypertension_date_b  0.0000000000
## ihd_date_b  0.0000000000
## Intellectual disability  0.0000000000
## isch_stroke_date_b  0.0000000000
## mood_anx_date_b  0.0000000000
## Mood disorders  0.0000000000
## ms_date_b  0.0000000000
## Malignant cell infestation  0.0000000000
## Neurocognitive disorders  0.0000000000
## osteoarthritis_date_b  0.0000000000
## osteoporosis_date_b  0.0000000000
## Osteomyelitis  0.0000000000
## parkinsonism_date_b  0.0000000000
## Personality disorders  0.0000000000
## Psychotic disorders  0.0000000000
## rheumatoid arthritis_date_b  0.0000000000
## Sedative and hypnotic use  0.0000000000
## Sepsis  0.0000000000
## Stimulant use  0.0000000000
## Tissue infestation  0.0000000000
### Tobaccouse 0.000000000
### total_od_n 0.000000000

#write.csv(rfevarimp, "rfevarimp.csv")

**Corrplots**

# library(corrplot)
#
# M <- cor(datasetmlmodel3)
#
# corrplot(M, method = "color")

# fatalodpop <- datasetmlmodel3 %>%
# filter(fatal_od_case == 1)
#
# library(corrplot)
#
# M1 <- cor(fatalodpop)
Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
Survival analysis code

```r
library(knitr)
library(dplyr)

## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##     filter, lag
```
## The following objects are masked from 'package:base':
##
## intersect, setdiff, setequal, union

library(survival)
library(ggplot2)
library(tibble)
library(caret)

## Loading required package: lattice

## Attaching package: 'caret'

## The following object is masked from 'package:survival':
##
## cluster

library(naniar)
library(ranger)
library(ggfortify)
library(readr)

#
# class(cleaneddata_v1$start_date)
# class(cleaneddata_v1$first_od_date)
#
# class(cleaneddata_v1$last_od_date)
#
# class(cleaneddata_v1$date_diff)
#
# cleaneddata_v1$date_diff <- difftime(cleaneddata_v1$last_od_date, cleaneddata_v1$start_date, units = c("days"))
#
# min(cleaneddata_v1$date_diff)
#
# max(cleaneddata_v1$date_diff)

#write.csv(cleaneddata_v2, "survivaldata_v1.csv", row.names = FALSE)

# cleaneddata_v2 <- select(cleaneddata_v1, -c(start_date, first_od_date, last_od_date, moh_study_id, od_case_id, to_keep))

survivaldata_v1 <- read_csv("U:/Andy's Thesis/MLoverdose/survivaldata_v1.csv", show_col_types = FALSE)

gg_miss_var(survivaldata_v1, show_pct = TRUE)

## Warning: The `guide` argument in `scale_*(*)` cannot be `FALSE`. This was deprecated in ggrepplot2 3.3.4.
## Please use "none" instead.
## The deprecated feature was likely used in the naniar package.
## Please report the issue at <https://github.com/njtierney/naniar/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
generated.

```r
km_fit <- survfit(Surv(date_diff, fatal_od_case) ~ 1, data = survivaldata_v1)
summary(km_fit, times = c(1, 30, 60, 90*(1:40)))
```

```r
## Call: survfit(formula = Surv(date_diff, fatal_od_case) ~ 1, data = survivaldata_v1)
##
## time n.risk n.event survival std.err lower 95% CI upper 95% CI
## 1 3 36659 8 1.000 7.71e-05 1.000 1.000
## 30 3 36399 50 0.998 2.08e-04 0.998 0.999
## 60 3 36161 57 0.997 2.94e-04 0.996 0.997
## 90 3 35909 38 0.996 3.39e-04 0.995 0.996
## 180 3 35000 156 0.991 4.86e-04 0.990 0.992
## 270 3 33986 179 0.986 6.17e-04 0.985 0.987
## 360 3 32958 215 0.980 7.48e-04 0.979 0.981
## 450 3 31654 259 0.972 8.87e-04 0.970 0.974
## 540 3 30401 212 0.966 9.90e-04 0.964 0.967
## 630 3 28937 247 0.958 1.11e-03 0.955 0.960
## 720 3 27021 383 0.944 1.28e-03 0.942 0.947
## 810 3 24883 454 0.928 1.47e-03 0.925 0.931
## 900 3 22432 465 0.910 1.66e-03 0.907 0.913
```
```{r}
 autoplot(km_fit)

 km_fit_sd <- survfit(Surv(date_diff, fatal_od_case) ~ Alcoholuse, data = survivaldata_v1)
 autoplot(km_fit_sd)
```
colnames(survivaldata_v1)

## [1] "fatal_od_case"
## [2] "total_od_n"
## [3] "alzheimer_dementia_date_b"
## [4] "ami_date_b"
## [5] "angina_date_b"
## [6] "asthma_date_b"
## [7] "ckd_date_b"
## [8] "copd_date_b"
## [9] "depression_date_b"
## [10] "diabetes_date_b"
## [11] "epilepsy_date_b"
## [12] "heart_failure_date_b"
## [13] "hypertension_date_b"
## [14] "hosp_stroke_date_b"
## [15] "haemorr_stroke_date_b"
## [16] "isch_stroke_date_b"
## [17] "hosp_tia_date_b"
## [18] "ihd_date_b"
## [19] "mood_anx_date_b"
## [20] "ms_date_b"
## [21] "osteo_arthritis_date_b"
## [22] "osteo_porosis_date_b"
## [23] "parkinsonism_date_b"
## [24] "rheumatoid_arthritis_date_b"
## [25] "oat"
## 

### [26] "substancerelateddisorders"
### [27] "Tobaccouse"
### [28] "Alcoholuse"
### [29] "Behaviouralpsychologicaldisturbances"
### [30] "Cannabinoiuse" 
### [31] "Cocaineuse"
### [32] "Developmentdisorders"
### [33] "earlyonsetdisorders"
### [34] "Endocarditis"
### [35] "Hallucinogensuse"
### [36] "Intellectualdisability"
### [37] "Mooddisorders"
### [38] "Multplementalillness"
### [39] "Neurocognitivedisorders"
### [40] "Neuroticrelateddisorders"
### [41] "Opioiduse"
### [42] "Osteomyelitis"
### [43] "Otherpsychoactivedruguse"
### [44] "Personalitydisorders"
### [45] "Polysubstance"
### [46] "psychoticdisorders"
### [47] "Sedativeandhypnoticuse"
### [48] "Sepsis"
### [49] "Stimulantuse"
### [50] "Tissueinfection"
### [51] "date_diff"

```r
cox <- coxph(Surv(date_diff, fatal_od_case) ~ total_od_n + alzheimer_dementia_date_b + ami_date_b + angina_date_b + asthma_date_b + ckd_date_b + copd_date_b + depression_date_b + diabetes_date_b + epilepsy_date_b + heart_failure_date_b + hypertension_date_b + hosp_stroke_date_b + haemorr_stroke_date_b + isch_stroke_date_b + hosp_tia_date_b + ihd_date_b + mood_anx_date_b + ms_date_b + osteoarthritis_date_b + osteoporosis_date_b + parkinsonism_date_b + rheumatoid_arthritis_date_b + oast + substancerelateddisorders + Tobaccouse + Alcoholuse + Behaviouralpsychologicaldisturbances + Cannabinoiuse + Cocaineuse + Developmentdisorders + earlyonsetdisorders + Endocarditis + Hallucinogensuse + Intellectualdisability + Mooddisorders + Multplementalillness + Neurocognitivedisorders + Neuroticrelateddisorders + Opioiduse + Osteomyelitis + Otherpsychoactivedruguse + Personalitydisorders + Polysubstance + psychoticdisorders + Sedativeandhypnoticuse + Sepsis + Stimulantuse + Tissueinfection, data = survivaldata_v1)
cox_fit <- survfit(cox)
autoplot(cox_fit)
```
aa_fit <- aareg(formula = Surv(date_diff, fatal_od_case) ~ total_od_n + alzheimer_dementia_date_b + ami_date_b + angina_date_b + asthma_date_b + ckd_date_b + copd_date_b + depression_date_b + diabetes_date_b + epilepsy_date_b + heart_failure_date_b + hypertension_date_b + hosp_stroke_date_b + haemorr_stroke_date_b + isch_stroke_date_b + hosp_tia_date_b + ihd_date_b + mood_anx_date_b + ms_date_b + osteo_arthritis_date_b + osteo_porosis_date_b + parkinsonism_date_b + rheumatoid_arthritis_date_b + oat + substancerelateddisorders + Tobaccouse + Alcoholuse + Behaviouralpsychologicaldisturbances + Cannabinoiduse + Cocaineuse + Developmentdisorders + earlyonsetdisorders + Endocarditis + Hallucinogensuse + Intellectualdisability + Mooddisorders + Multiplementalillness + Neurocognitivedisorders + Neuroticrelateddisorders + Opioiduse + Osteomyelitis + Otherpsychoactivedruguse + Personalitydisorders + Polysubstance + psychoticdisorders + Sedativeandhypnoticuse + Sepsis + Stimulantuse + Tissueinfection, data = survivaldata_v1)

aa_fit

## Call:
## aareg(formula = Surv(date_diff, fatal_od_case) ~ total_od_n +
##   alzheimer_dementia_date_b + ami_date_b + angina_date_b +
##   asthma_date_b + ckd_date_b + copd_date_b + depression_date_b +
##   diabetes_date_b + epilepsy_date_b + heart_failure_date_b +
##   hypertension_date_b + hosp_stroke_date_b + haemorr_stroke_date_b +
##   isch_stroke_date_b + hosp_tia_date_b + ihd_date_b + mood_anx_date_b +
##   ms_date_b + osteo_arthritis_date_b + osteo_porosis_date_b +
##   parkinsonism_date_b + rheumatoid_arthritis_date_b + oat +
##   substancerelateddisorders + Tobaccouse + Alcoholuse + Behaviouralpsychologicaldisturbances +
##   Cannabinoiduse + Cocaineuse + Developmentdisorders + earlyonsetdisorders +
##   Endocarditis + Hallucinogensuse + Intellectualdisability +
##   Mooddisorders + Multiplementalillness + Neurocognitivedisorders +
##   Opioiduse + Osteomyelitis + Otherpsychoactivedruguse + Personalitydisorders + Polysubstance + psychoticdisorders + Sedativeandhypnoticuse + Sepsis + Stimulantuse + Tissueinfection, data = survivaldata_v1)
## Neurocognitive disorders + Opioiduse + Osteomyelitis + Otherpsychoactivedruguse +
## Personalitydisorders + Polysubstance + psychotictdisorders +
## Sedativeandhypnoticuse + Sepsis + Stimulantuse + Tissueinfection,
## data = survivaldata_v1)
##
## n= 36679
## 1725 out of 1733 unique event times used
##
##         slope      coef   se(coef)    z
## Intercept 5.14e-04   8.99e-05  5.20e-06  17.3000
## total.od_n -4.87e-06 -1.35e-06  1.82e-07  -7.3700
## alzheimer_dementia_date_b -4.06e-05 -5.54e-06  5.96e-06  -0.9290
## ami_date_b  6.32e-05  9.64e-06  6.50e-06   1.4800
## angina_date_b -1.49e-05 -6.78e-06  6.10e-06  -1.1100
## asthma_date_b -1.98e-05 -3.92e-06  1.48e-06  -2.6500
## ckd_date_b  4.09e-05  1.04e-05  2.81e-06   3.6900
## copd_date_b  3.45e-05  8.22e-06  2.84e-06   2.8900
## depression_date_b  1.23e-04  2.10e-05  2.32e-06   9.0800
## diabetes_date_b -5.61e-06  5.18e-07  2.71e-06   0.1910
## epilepsy_date_b  8.77e-06  2.01e-06  3.36e-06   0.5990
## heart_failure_date_b  1.35e-04  2.21e-05  4.34e-06   5.1000
## hypertension_date_b  7.86e-05  1.23e-05  2.52e-06   4.8800
## hosp.stroke_date_b  4.26e-04  6.13e-05  3.99e-05   1.5400
## haemorr.stroke_date_b -3.65e-05  4.16e-06  3.71e-05  -0.1120
## isch.stroke_date_b -3.11e-04 -4.58e-05  3.93e-05  -1.1600
## hosp.tia_date_b -4.48e-05 -8.58e-06  1.27e-05  -0.6770
## ihd.date_b -2.13e-05 -2.11e-06  3.73e-06  -0.5660
## mood.anx_date_b -7.51e-06 -3.63e-06  2.48e-06  -1.4600
## ms.date_b  3.09e-05 -4.28e-06  1.26e-05  -0.3400
## osteo.arthritis.date_b  1.18e-05  2.61e-06  2.38e-06   1.0900
## osteo.porosis.date_b -6.91e-05 -1.23e-05  4.72e-06  -2.6200
## parkinsonism.date_b  1.81e-04  1.87e-05  1.74e-05   1.0800
## rheumatoid.arthritis.date_b -2.74e-05 -6.61e-06  5.04e-06  -1.3100
## oat -1.01e-04 -2.05e-05  1.32e-06  -15.5000
## substancerelateddisorders -1.17e-04 -1.89e-05  2.39e-06  -7.9000
## Tobacco -3.26e-05 -2.59e-06  2.40e-06  -1.0800
## Alcoholuse -2.72e-05 -1.57e-06  5.16e-06  -0.3040
## Behaviouralpsychologicaldisturbances -2.14e-05 -5.23e-06  4.63e-06  -1.1300
## Cannabinoiduse -8.14e-06 -9.76e-07  1.84e-06  -0.5310
## Cocaine -8.75e-05  1.67e-05  1.70e-06   9.8200
## Developmentdisorders -1.47e-05 -2.09e-06  3.17e-06  -0.6610
## earlyonsetdisorders -6.18e-05 -9.72e-06  1.44e-06  -6.7700
## Endocarditis  1.27e-05 -2.21e-07  3.28e-06  -0.0674
## Hallucinogenusenuse -5.41e-05 -1.08e-05  4.50e-06  -2.4100
## Intellectualdisability -4.19e-05 -1.02e-05  4.88e-06  -2.0800
## Mooddisorders -3.06e-05 -4.27e-06  1.82e-06  -2.3500
## Multiplémentailliness -2.40e-05 -4.41e-06  1.46e-06  -3.0200
## Neurocognitivedisorders -2.91e-05 -4.58e-06  1.83e-06  -2.5000
## Neuroticrelateddisorders -9.17e-05 -1.52e-05  1.50e-06  -10.1000
## Opioiduse -5.24e-05 -1.15e-05  1.51e-06   -7.6300
## Osteomyelitis -1.85e-05 -5.85e-06  2.76e-06  -2.1200
<table>
<thead>
<tr>
<th>Category</th>
<th>Coefficient</th>
<th>95% Confidence Interval</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other psychoactive drug use</td>
<td>-6.54e-05</td>
<td>-1.38e-05 1.47e-06</td>
<td>9.3800</td>
</tr>
<tr>
<td>Personality disorders</td>
<td>1.81e-05</td>
<td>2.47e-06 1.70e-06</td>
<td>1.4600</td>
</tr>
<tr>
<td>Polysubstance</td>
<td>-2.24e-05</td>
<td>-5.10e-06 1.70e-06</td>
<td>3.0100</td>
</tr>
<tr>
<td>Psychotic disorders</td>
<td>-1.52e-05</td>
<td>-1.92e-06 1.68e-06</td>
<td>1.1400</td>
</tr>
<tr>
<td>Sedative and hypnotic use</td>
<td>3.87e-05</td>
<td>7.06e-06 2.97e-06</td>
<td>2.3800</td>
</tr>
<tr>
<td>Sepsis</td>
<td>-2.70e-05</td>
<td>-3.65e-06 1.81e-06</td>
<td>2.0200</td>
</tr>
<tr>
<td>Stimulant use</td>
<td>-4.14e-05</td>
<td>-8.03e-06 1.67e-06</td>
<td>4.8100</td>
</tr>
<tr>
<td>Tissue infection</td>
<td>-4.23e-05</td>
<td>-7.25e-06 1.55e-06</td>
<td>4.6700</td>
</tr>
<tr>
<td>Intercept</td>
<td>5.38e-67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total OD N</td>
<td>1.66e-13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alzheimer dementia date b</td>
<td>3.53e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ami date b</td>
<td>1.38e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angina date b</td>
<td>2.67e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asthma date b</td>
<td>8.07e-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CKD date b</td>
<td>2.22e-04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COPD date b</td>
<td>3.82e-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression date b</td>
<td>1.07e-19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes date b</td>
<td>8.48e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epilepsy date b</td>
<td>5.49e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heart failure date b</td>
<td>3.43e-07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypertension date b</td>
<td>1.07e-06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hosp stroke date b</td>
<td>1.24e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haemorrh stroke date b</td>
<td>9.11e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isch stroke date b</td>
<td>2.44e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hosp tia date b</td>
<td>4.98e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IHF date b</td>
<td>5.71e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood anx date b</td>
<td>1.43e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS date b</td>
<td>7.34e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Osteo arthritis date b</td>
<td>2.74e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Osteo porosis date b</td>
<td>8.86e-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parkinsonism date b</td>
<td>2.81e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rheumatoid arthritis date b</td>
<td>1.90e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OAT</td>
<td>2.75e-54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Substance related disorders</td>
<td>2.81e-15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tobacco</td>
<td>2.80e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alcohol use</td>
<td>7.61e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behavioural psychological disturbances</td>
<td>2.58e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cannabinoid use</td>
<td>5.95e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cocaine use</td>
<td>9.11e-23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Development disorders</td>
<td>5.08e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early onset disorders</td>
<td>1.30e-11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endocarditis</td>
<td>9.46e-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hallucinogen use</td>
<td>1.61e-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intellectual disability</td>
<td>3.73e-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood disorders</td>
<td>1.90e-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiplemantillness</td>
<td>2.55e-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neurocognitive disorders</td>
<td>1.25e-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neurotic related disorders</td>
<td>5.09e-24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Opioid use</td>
<td>2.27e-14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Osteomyelitis</td>
<td>3.42e-02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
##  Other psychoactive drug use             6.53e-21
##  Personality disorders                 1.46e-01
##  Polysubstance                          2.65e-03
##  Psychotic disorders                    2.53e-01
##  Sedative and hypnotic use              1.73e-02
##  Sepsis                                  4.34e-02
##  Stimulant use                           1.50e-06
##  Tissue infection                        3.04e-06
##
##  Chisq = 2286.35 on 49 df, p = <2e-16; test weights = aalen

```r
plot <- autoplot(aa_fit)
```

---

---

```r
# Warning: `gather_()` was deprecated in tidyr 1.2.0.
# i Please use `gather()` instead.
# i The deprecated feature was likely used in the ggfortify package.
# Please report the issue at <https://github.com/sinhrks/ggfortify/issues>.
# This warning is displayed once every 8 hours.
# Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
# generated.

# Warning: `mutate_()` was deprecated in dplyr 0.7.0.
# i Please use `mutate()` instead.
# i See vignette('programming') for more help
# i The deprecated feature was likely used in the ggfortify package.
# Please report the issue at <https://github.com/sinhrks/ggfortify/issues>.
# This warning is displayed once every 8 hours.
# Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
# generated.

# Warning: `group_by_()` was deprecated in dplyr 0.7.0.
# i Please use `group_by()` instead.
# i See vignette('programming') for more help
# i The deprecated feature was likely used in the ggfortify package.
# Please report the issue at <https://github.com/sinhrks/ggfortify/issues>.
# This warning is displayed once every 8 hours.
# Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
# generated.

plot
r_fit <- ranger(Surv(date_diff, fatal_od_case) ~ total_od_n + alzheimer_dementia_date_b + ami_date_b + angina_date_b + asthma_date_b + ckd_date_b + copd_date_b + depression_date_b + diabetes_date_b + epilepsy_date_b + heart_failure_date_b + hypertension_date_b + hosp_stroke_date_b + haemorr_stroke_date_b + isch_stroke_date_b + hosp_tia_date_b + ihd_date_b + mood_anx_date_b + ms_date_b + osteo_arthritis_date_b + osteo_porosis_date_b + parkinsonism_date_b + rheumatoid_arthritis_date_b + oot + substance-relateddisorders + Tobacco + Alcoholuse + Behaviouralpsychologicaldisturbances + Cannabinoiduse + Cocaineuse + Developmentdisorders + earlyonsetdisorders + Endocarditis + Hallucinogensuse + Intellectualdisability + Mooddisorders + Multiplementalliness + Neurocognitive disorders + Neuroticrelateddisorders + Opioiduse + Osteomyelitis + Otherpsychoactive druguse + Personalitydisorders + Polysubstance + psychoticdisorders + Sedativeandhypnoticuse + Sepsis + Stimulantuse + Tissueinfection,

data = survivaldata_v1,
mtry = 4,
importance = "permutation",
splitrule = "extratrees",
verbose = TRUE)

## Computing permutation importance.. Progress: 0%. Estimated remaining time: 4 hours, 17 minutes, 49 seconds.
death_times <- r_fit$unique.death.times
surv_prob <- data.frame(r_fit$survival)
avg_prob <- sapply(surv_prob, mean)

plot(r_fit$unique.death.times, r_fit$survival[1,],
     type = "l",
     ylim = c(0,1),
     col = "red",
     xlab = "days",
     ylab = "survival",
     main = "Patient Survival Curves")

cols <- colors()

for(n in sample(c(2:dim(survivaldata_v1)[1]), 20)){
  lines(r_fit$unique.death.times, r_fit$survival[n,], type = "l", col = cols[2])
}
lines(death_times, avg_prob, lwd = 2)
legend(500, 0.7, legend = c('Average = black'))

vi <- data.frame(sort(round(r_fit$variable.importance, 4), decreasing = TRUE))

names(vi) <- "importance"
head(vi)

##                           importance
## Opioiduse                     0.0328
## oat                           0.0190

vi
## Polysubstance                 0.0165
## substance-related disorders   0.0134
## Neurotic-related disorders    0.0123
## Other psychoactive drug use   0.0105

```
> cat("Prediction Error = 1 - Harrell's c-Index = 1 + ROC =", r_fit$prediction.error)
## Prediction Error = 1 - Harrell's c-Index = 1 + ROC = 0.3253922
```

```r
kmi <- rep("KM", length(km_fit$time))
k_km_df <- data.frame(km_fit$time, km_fit$surv, kmi)
names(k_km_df) <- c("Time", "surv", "Model")

coxi <- rep("Cox", length(cox_fit$time))
c_cox_df <- data.frame(cox_fit$time, cox_fit$surv, coxi)
names(c_cox_df) <- c("Time", "surv", "Model")

rfi <- rep("RF", length(r_fit$unique.death.times))
r_rf_df <- data.frame(r_fit$unique.death.times, avg_prob, rfi)
names(r_rf_df) <- c("Time", "surv", "Model")

plot_df <- rbind(k_km_df, c_cox_df, r_rf_df)

p <- ggplot(plot_df, aes(x = Time, y = surv, color = Model))
p+geom_line()
```
\texttt{saveRDS(km\_fit, file="km\_fit.rds")}

\texttt{saveRDS(cox\_fit, file="cox\_fit.rds")}

\texttt{saveRDS(r\_fit, file="r\_fit.rds")}
D.5  ML Prediction Code for Overdose

Data preprocessing

```r
library(tidyverse)
library(dplyr)
library(gapminder)
library(caret)
library(xgboost)
library(forcats)
library(mice)
library(vctrs)
library(naniar)
library(arsenal)
library(janitor)

##loading data
ICDderivedvariables <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/og data/ICDderivedvariables.csv", show_col_types = FALSE)
Pharmanetdinpin <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/og data/Pharmanetdinpin.csv", show_col_types = FALSE)
vw_cdr_case_b <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/og data/vw_cdr_case_b.csv", show_col_types = FALSE)
vw_od_case <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/og data/vw_od_case.csv", show_col_types = FALSE)

glimpse(ICDderivedvariables)
glimpse(Pharmanetdinpin)
```

## Rows: 57,902,314
## Columns: 3

```r
length(unique(vw_od_case$moh_study_id)) == nrow(vw_od_case)
```

## [1] TRUE

```r
n_distinct(ICDderivedvariables$moh_study_id)
```

## [1] 1094483

```r
n_distinct(Pharmanetdinpin$moh_study_id)
```

## [1] 647873

```r
n_distinct(vw_cdr_case_b$moh_study_id)
```
```r
## 1] 608182

# dataset1 <- left_join(vw_cdr_case_b, Pharmanetdinpin, by = "moh_study_id")
#
# controls <- left_join(dataset1, ICDderivedvariables, by = "moh_study_id")
#
# length(unique(controls$moh_study_id)) == nrow(controls)
#
# controls1 <- controls %>%
# distinct(moh_study_id, .keep_all = TRUE)
#
# length(unique(controls1$moh_study_id)) == nrow(controls1)
#
# glimpse(controls1)
#
# controls2 <- controls1 %>%
# select(moh_study_id, ost, rheumatoid_arthritis_date_b, parkinsonism_date_b, osteo_arthritis_date_b,
# osteo_porosis_date_b, ms_date_b, mood_anx_date_b, isch_stroke_date_b, ihd_date_b, hypertension_date_b,
# hosp_ita_date_b, hosp_stroke_date_b, heart_failure_date_b, haemorrh_stroke_date_b, epilepsy_date_b,
# diabetes_date_b, depression_date_b, copd_date_b, ckd_date_b, asthma_date_b, angina_date_b, ami_date_b,
# alzheimer_dementia_date_b, Tobaccouse, Tissueinfection, substancerelateddisorders.x, Stimulantuse,
# Sepsis, Sedativeandhypnoticuse, psychoticdisorders, Polysubstance, Personalitydisorders, Opioiduse,
# Neurocognitivedisorders, Multiplementalillness, Mooddisorders, Intellectualdisability, Hallucinogensuse,
# Endocarditis, earlyonsetdisorders, Developmentdisorders, Cocaineuse, Cannabinoiduse, Behaviouralpsychologicaldisturbances, Alcoholuse)
#
# controls3 <- na.omit(controls2)
#
# gg_miss_var(controls1)
#
# gg_miss_var(controls3)
#
# odmohs <- vw_od_case %>%
# select(moh_study_id)

# controls4 <- subset(controls3, !(moh_study_id %in% odmohs$moh_study_id))

# controls5 <- sample_n(controls4, 36679)
#
# controls6 <- controls5 %>%
# select(-moh_study_id) %>%
# mutate(od_case = 0)

# write.csv(controls4, "controlsbigpool.csv", row.names = FALSE)
# write.csv(controls6, "controls.csv", row.names = FALSE)

controlsbigpool <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/controlsbigpool.csv", show_col_types = FALSE)

glimpse(controlsbigpool)
```
dataset2 <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/og data/datasetfirst.csv", show_col_types = FALSE)

controlsbinpool <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/controlsbigpool.csv", show_col_types = FALSE)

glimpse(controlsbinpool)

# controls6 <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/controls.csv", show_col_types = FALSE)

dataset3 <- dataset2 %>%
  distinct(moh_study_id, .keep_all = TRUE)

gg_miss_var(dataset3, show_pct = TRUE)
glimpse(dataset3)

## Rows: 36,679
## Columns: 67

na_count <- sapply(dataset3, function(y) sum(length(which(is.na(y)))))

na_count <- data.frame(na_count)

data1 <- dataset3 %>%
  select(moh_study_id, ost, rheumatoid_arthritis_date_b, parkinsonism_date_b, osteo_arthritis_date_b, osteo_porosis_date_b, ms_date_b, mood_anx_date_b, isch_stroke_date_b, ihd_date_b, hypertension_date_b, hosp_tia_date_b, hosp_stroke_date_b, heart_failure_date_b, haemorr_stroke_date_b, epilepsy_date_b, diabetes_date_b, depression_date_b, ckd_date_b, asthma_date_b, angina_date_b, ami_date_b, alzheimer_dementia_date_b, Tobaccouse, Tissueinfection, substancerelateddisorders.x, Stimulantuse, Sepsis, Sedativeandhypnoticuse, psychoticdisorders, Polysubstance, Personalitydisorders, Otherpsychoactivedruguse, Osteomyelitis, Opioiduse, Neurocognitivedisorders, Multiplementalillness, Mooddisorders, Intellectualdisability, Hallucinogensuse, Cardiacdisorders, Developmentaldisorders, Cerebrovasculardisease, Behavourspsychologicaldisturbances, Alcoholuse, total_dr, tcalc_2, hba1c, moh_study_id, last_og_date, first_og_date, total_og_date)
opmentdisorders, Cocaineuse, Cannabinoiduse, Behaviouralpsychologicaldisturbances, Alcoholuse)%>
%
mutate(od_case = 1)
data2 <- data1[,-1]
controlsbigpool1 <- controlsbigpool %>%
select(-moh_study_id)%>%
mutate(od_case = 0)

compare_df_cols(data2, controlsbigpool1)

##                             column_name   data2 controlsbigpool1
## 1                            Alcoholuse numeric          numeric
## 2             alzheimer_dementia_date_b numeric          numeric
## 3                            ami_date_b numeric          numeric
## 4                         angina_date_b numeric          numeric
## 5                         asthma_date_b numeric          numeric
## 6  Behaviouralpsychologicaldisturbances numeric          numeric
## 7                        Cannabinoiduse numeric          numeric
## 8                            ckd_date_b numeric          numeric
## 9                            Cocaineuse numeric          numeric
## 10                        copd_date_b numeric          numeric
## 11                    depression_date_b numeric          numeric
## 12                    Developmentdisorders numeric          numeric
## 13                         diabetes_date_b numeric          numeric
## 14                  earlyonsetdisorders numeric          numeric
## 15                      Endocarditis numeric          numeric
## 16                epilepsy_date_b numeric          numeric
## 17                  haemorr_stroke_date_b numeric          numeric
## 18                   Hallucinogensuse numeric          numeric
## 19                 heart_failure_date_b numeric          numeric
## 20                  hosp_stroke_date_b numeric          numeric
## 21                    hosp_tia_date_b numeric          numeric
## 22               hypertension_date_b numeric          numeric
## 23                        ihd_date_b numeric          numeric
## 24                Intellectualdisability numeric          numeric
## 25                isch_stroke_date_b numeric          numeric
## 26                     mood_anx_date_b numeric          numeric
## 27                          Mooddisorders numeric          numeric
## 28                         ms_date_b numeric          numeric
## 29              Multiplementalillness numeric          numeric
## 30                Neurocognitivedisorders numeric          numeric
## 31                  Neuroticrelateddisorders numeric          numeric
## 32                            od_case numeric          numeric
## 33                            Opioiduse numeric          numeric
## 34                              ost numeric          numeric
## 35                osteo_arthritis_date_b numeric          numeric
```r
## 36     osteo_porosis_date_b numeric numeric
## 37       Osteomyelitis numeric numeric
## 38     Otherpsychoactive druguse numeric numeric
## 39  parkinsonism_date_b numeric numeric
## 40     Personalitydisorders numeric numeric
## 41      Polysubstance numeric numeric
## 42   psychoticdisorders numeric numeric
## 43  rhematoid_arthritis_date_b numeric numeric
## 44  Sedativeandhypnoticuse numeric numeric
## 45           Sepsis numeric numeric
## 46        Stimulantuse numeric numeric
## 47 substancerelateddisorders.x numeric numeric
## 48         Tissueinfection numeric numeric
## 49       Tobaccouse numeric numeric

overdosedata <- rbind(data2, controlsbigpool1)

table(overdosedata$od_case)

##
##   0   1
## 446710 36679

# inferencedataset1 <- as.data.frame(unclass(overdosedata))
#
# inferencedataset1 <- data.frame(complete(mice(inferencedataset1, m=7, method = "cart", seed= 123, remove.collinear=FALSE)))
#
# write.csv(inferencedataset1, "inferencedataset1.csv", row.names = FALSE)

### splitting before imputation

# set.seed(2001)
#
# data <- overdosedata
#
# index= createDataPartition(y=data$od_case, p = 0.7, list = FALSE)
#
# train.set = data[index,]
# test.set = data[-index,]
#
# test.set$od_case = as.factor(test.set $od_case)
# train.setSod_case = as.factor(train.setSod_case)
#
# train.set = train.set %>% mutate_if(is.character, as.factor)
# test.set = test.set %>% mutate_if(is.character, as.factor)
#
# dim(train.set)
# dim(test.set)
```

455
# table(train.set$od_case)
# table(test.set$od_case)
#
# na_count <- sapply(train.set, function(y) sum(length(which(is.na(y)))))
# na_count1 <- data.frame(na_count)
# na_count1
#
# na_count2 <- sapply(test.set, function(y) sum(length(which(is.na(y)))))
# na_count3 <- data.frame(na_count2)
# na_count3

train.setna <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/train.setna.csv", show_col_types = FALSE)

test.setna <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/test.setna.csv", show_col_types = FALSE)

na_count0 <- sapply(train.setna, function(y) sum(length(which(is.na(y)))))

na_counttrain <- data.frame(na_count0)

na_counttrain
##                                        na_count0
## ost                                      2282
## rheumatoid_arthritis_date_b            3612
## parkinsonism_date_b                     3612
## osteo_arthritis_date_b                  3612
## osteo_porosis_date_b                    3612
## ms_date_b                               3612
## mood_anx_date_b                         3612
## isch_stroke_date_b                      3612
## ihd_date_b                              3612
## hypertension_date_b                     3612
## hosp_tia_date_b                         3612
## hosp_stroke_date_b                      3612
## heart_failure_date_b                    3612
## haemorr_stroke_date_b                   3612
## epilepsy_date_b                         3612
## diabetes_date_b                         3612
## depression_date_b                       3612
## copd_date_b                             3612
## ckd_date_b                              3612
## asthma_date_b                           3612
## angina_date_b                           3612
na_count1 <- sapply(test.setna, function(y) sum(length(which(is.na(y)))))

na_counttest <- data.frame(na_count1)

na_counttest

##                        na_count1
## ost                      1024
## rheumatoid_arthritis_date_b 1563
## parkinsonism_date_b     1563
## osteo_arthritis_date_b   1563
## osteo_porosis_date_b    1563
## ms_date_b               1563
## mood_anx_date_b         1563
## isch_stroke_date_b      1563
## ihd_date_b              1563
## hypertension_date_b     1563
## hosp_tia_date_b         1563
## hosp_stroke_date_b      1563
## heart_failure_date_b    1563
## haemorr_stroke_date_b   1563
## epilepsy_date_b         1563
```r
## diabetes_date_b  1563
## depression_date_b  1563
## copd_date_b  1563
## ckd_date_b  1563
## asthma_date_b  1563
## angina_date_b  1563
## ami_date_b  1563
## alzheimer_dementia_date_b  1563
## Tobaccouse  533
## Tissueinfection  533
## substancerelated.disorders  533
## Stimulantuse  533
## Sepsis  533
## Sedativeandhypnoticuse  533
## psychoticdisorders  533
## Polysubstance  533
## Personalitydisorders  533
## Otherpsychoactivedruguse  533
## Osteomyelitis  533
## Opioiduse  533
## Neuroticrelateddisorders  533
## Neurocognitivedisorders  533
## Multiplementalillness  533
## Mooddisorders  533
## Intellectualdisability  533
## Hallucinogensuse  533
## Endocarditis  533
## earlyonsetdisorders  533
## Developmentdisorders  533
## Cocaineuse  533
## Cannabinoiduse  533
## Behaviouralpsychologicaldisturbances  533
## Alcoholuse  533
## od_case  0

# set2 <- train.set[sample(which(train.set$od_case != "1"), 25671), ]
#
# table(set2$od_case)
#
# set1 <- train.set %>%
#   filter(od_case == 1)
#
# train.set1 <- rbind(set2, set1)
#
# table(train.set$od_case)
# table(train.set1$od_case)
# table(test.set$od_case)

# test.set1 <- test.set
```

458
```r
train.set1 <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/train.set2.csv", show_column_types = FALSE)

test.set1 <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/test.set2.csv", show_column_types = FALSE)

gg_miss_var(train.set1, show_pct = TRUE)

glimpse(train.set1)
## Rows: 51,342
## Columns: 49

table(train.set1$od_case)
```
gg_miss_var(test.set1, show_pct = TRUE)

```r
gg_miss_var(test.set1, show_pct = TRUE)
```

glimpse(test.set1)

```r
glimpse(test.set1)
```

```r
## Rows: 145,016
## Columns: 49
table(test.set1$od_case)

```r
table(test.set1$od_case)

```r
## 0 1
## 134008 11008

```r
# train.set2 <- train.set1 %>%
# rename(substancerelated.disorders = substancerelateddisorders.x)
```
# test.set2 <- test.set1 %>%
# rename(substancerelated.disorders = substancerelateddisorders.x)

# write.csv(train.set2, "train.setna.csv", row.names = FALSE)
# write.csv(test.set2, "test.setna.csv", row.names = FALSE)

# train.set3 <- as.data.frame((unclass(train.set2)))
#
# train.set3 <- data.frame(complete(mice(train.set3, m=7, method = "cart", seed= 123, remove.collinear = FALSE)))
#
# test.set3 <- as.data.frame((unclass(test.set2)))
#
# test.set3 <- data.frame(complete(mice(test.set3, m=7, method = "cart", seed= 123, remove.collinear = FALSE)))
#
# write.csv(train.set3, "train.set2.csv", row.names = FALSE)
# write.csv(test.set3, "test.set2.csv", row.names = FALSE)

ML modeling with Caret

library(keras)
library(tensorflow)
library(tidyverse)
library(dplyr)
library(gapminder)
library(caret)
library(xgboost)
library(forcats)
library(mice)
library(vctrs)
library(naniar)
library(yardstick)
library(forcats)
library(Boruta)
library(ROSE)
library(imbalance)
library(rnn)
library(ROCR)
library(SHAPforxgboost)
library(randomForest)
library(mlbench)
library(ggplot2)
library(data.table)
library(here)
library(pROC)
library(ROCR)
### Testing the original sample

dataset_train <- read_csv("U:/Andy's Thesis/ML overdose in general/GeneralodML/train.set2.csv", show_col_types = FALSE)

dataset_test <- read_csv("U:/Andy's Thesis/ML overdose in general/GeneralodML/test.set2.csv", show_col_types = FALSE)

train.set <- dataset_train
test.set <- dataset_test

train.set <- train.set %>%
  mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))

test.set <- test.set %>%
  mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))

train.set = train.set %>% mutate_if(is.character, as.factor)
test.set = test.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 51342  49

dim(test.set)
## [1] 145016  49

nrow(train.set)
## [1] 51342

table(train.set$od_case)
##
##    No  Yes
## 25671 25671

dim(train.set)
## [1] 51342  49

dim(test.set)
## [1] 145016  49

table(test.set$od_case)
##
##    No  Yes
## 25671 25671

table(test.set$od_case)
```r
##
## No    Yes
## 134008 11008

class(train.set$Sod_case)
## [1] "factor"
class(test.set$Sod_case)
## [1] "factor"

#XGBOOST

grid_default <- expand.grid(
  nrounds= 300,
  max_depth = 5,
  eta = 0.3, gamma = 0.001,
  colsample_bytree = 0.75,
  min_child_weight = 1,
  subsample = 1
)

train_control <- caret::trainControl(
  method = "none",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
  allowParallel = TRUE
)

model = caret::train(od_case ~ .,
  data = train.set,
  trControl = train_control,
  method = "xgbTree",
  tuneGrid = grid_default,
  verbose = TRUE
)

pred <- predict(model, test.set)

confusionMatrix(factor(pred), factor(test.set$Sod_case), positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference Prediction  No  Yes
## No  119364  2006
## Yes  14644  9002
##
##          Accuracy : 0.8852
## 95% CI : (0.8835, 0.8868)
```
## No Information Rate : 0.9241
## P-Value [Acc > NIR] : 1
##
## Kappa : 0.464
##
## McNemar's Test P-Value : <2e-16
##
## Sensitivity : 0.81777
## Specificity : 0.89072
## Pos Pred Value : 0.38070
## Neg Pred Value : 0.98347
## Prevalence : 0.07591
## Detection Rate : 0.06208
## Detection Prevalence : 0.16306
## Balanced Accuracy : 0.85425
##
## 'Positive' Class : Yes
##
prob <- predict(model, test.set, type = "prob")
results.roc <- roc(test.set$od_case, prob$Yes)
auc(results.roc)

## Area under the curve: 0.9084

plot(results.roc, print.thres = "best", print.thres.best.method = "closest.topleft")
results.coords <- coords(results.roc,"best", best.method = "closest.topleft")
print(results.coords)

##  threshold specificity sensitivity
## 1 0.4577166   0.8824249   0.8254906

# Random Forest

grid_default <- expand.grid(
  .mtry = 1:7,
  .splitrule="gini",
  .min.node.size=c(1)
)

train_control <- caret::trainControl(
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
  allowParallel = TRUE
)

rf2 = caret::train(od_case ~ .,
  data = train.set,
  method = "ranger",
  trControl = train_control,
  tuneGrid = grid_default,
  verbose = TRUE
)

pred <- predict(rf2, test.set)
confusionMatrix(factor(pred), factor(test.set$od_case), positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference
## Prediction No Yes
## No 120098 2047
## Yes 13910 8961
##
## Accuracy : 0.89
## 95% CI : (0.8883, 0.8916)
## No Information Rate : 0.9241
## P-Value [Acc > NIR] : 1
##
## Kappa : 0.4752
##
## Mcnemar's Test P-Value : <2e-16
##
## Sensitivity : 0.81404
## Specificity : 0.89620
##          Pos Pred Value : 0.39181
##          Neg Pred Value : 0.98324
##          Prevalence : 0.07591
##          Detection Rate : 0.06179
##    Detection Prevalence : 0.15771
##          Balanced Accuracy : 0.85512
##
##        'Positive' Class : Yes
##
prob <- stats::predict(rf2, newdata = test.set, type = "prob")

results.roc <- roc(test.set$Sod_case, prob$Yes)
auc(results.roc)

## Area under the curve: 0.9111

plot(results.roc, print.thres = "best", print.thres.best.method = "closest.topleft")

results.coords <- coords(results.roc,"best", best.method = "closest.topleft")
print(results.coords)

##  threshold specificity sensitivity
## 1 0.4404112   0.8814623    0.828125

###feature selected (From Boruta)
datasettrain <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/train.set2.csv", show_col_types = FALSE)

datasettest <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/test.set2.csv", show_col_types = FALSE)

train.set <- datasettrain
test.set <- datasettest

train.set <- train.set %>%
  mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))

test.set <- test.set %>%
  mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))

train.set = train.set %>% mutate_if(is.character, as.factor)
test.set = test.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 51342  49
dim(test.set)
## [1] 145016  49

nrow(train.set)
## [1] 51342

table(train.set$od_case)
##
## No   Yes
## 25671 25671
dim(train.set)
## [1] 51342  49
dim(test.set)
## [1] 145016  49
table(test.set$od_case)
##
## No   Yes
## 25671 25671

table(test.set$od_case)
##  No  Yes
## 134008 11008

```r
class(train.set$od_case)
## [1] "factor"

class(test.set$od_case)
## [1] "factor"
```

```r
train.set <- train.set%>%
  select(od_case, ost, rheumatoid_arthritis_date_b, parkinsonism_date_b, osteo_arthritis_date_b, osteo_porosis_date_b, mood_anx_date_b, isch_stroke_date_b, ihd_date_b, hypertension_date_b, hosp_stroke_date_b, heart_failure_date_b, haemorr_stroke_date_b, epilepsy_date_b, diabetes_date_b, depression_date_b, copd_date_b, ckd_date_b, asthma_date_b, angina_date_b, ami_date_b, alzheimer_dementia_date_b, Tobacco, Tissueinfection, substancerelated.disorders, Stimulantuse, Sepsis, Sedativeandhypnoticuse, psychoticdisorders, Polysubstance, Personalitydisorders, Otherpsychoactivedruguse, Osteomyelitis, Opioiduse, Neuroticatedisorders, Multisubstance, Mooddisorders, Intellectualdisability, Hallucinogenicuse, Endocarditis, earlyonsetdisorders, Developmentdisorders, Cocaineuse, Cannabinoiduse, Behaviouralpsychologicaldisturbances, Alcoholuse)
```

```r
test.set <- test.set%>%
  select(od_case, ost, rheumatoid_arthritis_date_b, parkinsonism_date_b, osteo_arthritis_date_b, osteo_porosis_date_b, mood_anx_date_b, isch_stroke_date_b, ihd_date_b, hypertension_date_b, hosp_stroke_date_b, heart_failure_date_b, haemorr_stroke_date_b, epilepsy_date_b, diabetes_date_b, depression_date_b, copd_date_b, ckd_date_b, asthma_date_b, angina_date_b, ami_date_b, alzheimer_dementia_date_b, Tobacco, Tissueinfection, substancerelated.disorders, Stimulantuse, Sepsis, Sedativeandhypnoticuse, psychoticdisorders, Polysubstance, Personalitydisorders, Otherpsychoactivedruguse, Osteomyelitis, Opioiduse, Neuroticatedisorders, Multisubstance, Mooddisorders, Intellectualdisability, Hallucinogenicuse, Endocarditis, earlyonsetdisorders, Developmentdisorders, Cocaineuse, Cannabinoiduse, Behaviouralpsychologicaldisturbances, Alcoholuse)
```

```r
grid_default <- expand.grid(
  nrounds = 300,
  max_depth = 5,
  eta = 0.3, gamma = 0.001,
  colsample_bytree = 0.75,
  min_child_weight = 1,
  subsample = 1
)
```

```r
train_control <- caret::trainControl(
  method = "none",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
)
allowParallel = TRUE

model = caret::train(od_case ~ .,
    data = train.set,
    trControl = train_control,
    method = "xgbTree",
    tuneGrid = grid_default,
    verbose = TRUE)

pred <- predict(model, test.set)

confusionMatrix(factor(pred), factor(test.set$sod_case), positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference  No  Yes
## Prediction
##     No  119375  2015
##     Yes  14633  8993
##
##                Accuracy : 0.8852
##          95% CI : (0.8835, 0.8868)
## No Information Rate : 0.9241
## P-Value [Acc > NIR] : 1
##
##          Kappa : 0.4638
##
## McNemar's Test P-Value : <2e-16
##
## Sensitivity : 0.81695
## Specificity : 0.89081
## Pos Pred Value : 0.38064
## Neg Pred Value : 0.98340
## Prevalence : 0.07591
## Detection Rate : 0.06201
## Detection Prevalence : 0.16292
## Balanced Accuracy : 0.85388
##
## 'Positive' Class : Yes
##
prob <- predict(model, test.set, type = "prob")
results.roc <- roc(test.set$sod_case, prob$Yes)
auc(results.roc)

## Area under the curve: 0.9083

plot(results.roc, print.thres = "best", print.thres.best.method = "closest.topleft")
```r
results.coords <- coords(results.roc, "best", best.method = "closest.topleft")
print(results.coords)
## threshold specificity sensitivity
## 1 0.4511632 0.881522 0.8253089

datasettrain <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/train.set2.csv", show_col_types = FALSE)
datasettest <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/test.set2.csv", show_col_types = FALSE)

train.set <- datasettrain
test.set <- datasettest

train.set <- train.set %>%
mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))
test.set <- test.set %>%
mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))

train.set = train.set %>% mutate_if(is.character, as.factor)
test.set = test.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 51342 49
```
train.set <- train.set %>%
  select(od_case, ost, rheumatoid_arthritis_date_b, parkinsonism_date_b, osteo_arthritis_date_b, osteo_poro
  resis_date_b, mood_anx_date_b, isch_stroke_date_b, ihd_date_b, hypertension_date_b, hosp_stroke_date
  _b, heart_failure_date_b, haemorr_stroke_date_b, epilepsy_date_b, diabetes_date_b, depression_date_b, c
  opd_date_b, ckd_date_b, asthma_date_b, angina_date_b, ami_date_b)

test.set <- test.set %>%
  select(od_case, ost, rheumatoid_arthritis_date_b, parkinsonism_date_b, osteo_arthritis_date_b, osteo_poro
  resis_date_b, mood_anx_date_b, isch_stroke_date_b, ihd_date_b, hypertension_date_b, hosp_stroke_date
  _b, heart_failure_date_b, haemorr_stroke_date_b, epilepsy_date_b, diabetes_date_b, depression_date_b, c
  opd_date_b, ckd_date_b, asthma_date_b, angina_date_b, ami_date_b)
grid_default <- expand.grid(
  nrounds= 300,
  max_depth = 5,
  eta = 0.3, gamma = 0.001,
  colsample_bytree = 0.75,
  min_child_weight = 1,
  subsample = 1
)

train_control <- caret::trainControl(
  method = "none",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
  allowParallel = TRUE
)

model = caret::train(od_case ~ .,
  data = train.set,
  trControl = train_control,
  method = "xgbTree",
  tuneGrid = grid_default,
  verbose = TRUE
)

pred <- predict(model, test.set)

confusionMatrix(factor(pred), factor(test.set$sod_case), positive = "Yes")

## Confusion Matrix and Statistics
## Reference
## Prediction  No  Yes
##      No  86027  3164
##      Yes 47981  7844

## accuracy : 0.6473
## 95% CI : (0.6448, 0.6498)
## No Information Rate : 0.9241
## P-Value [Acc > NIR] : 1
##
## Kappa : 0.1236

## McNemar's Test P-Value : <2e-16
##
## Sensitivity : 0.71257
## Specificity : 0.64195
## Pos Pred Value : 0.14051
## Neg Pred Value : 0.96453
## Prevalence : 0.07591
## Detection Rate: 0.05409
## Detection Prevalence: 0.38496
## Balanced Accuracy: 0.67726
##
## 'Positive' Class: Yes
##
prob <- predict(model, test.set, type = "prob")
results.roc <- roc(test.set$od_case, prob$Yes)
auc(results.roc)

## Area under the curve: 0.7611
plot(results.roc, print.thres = "best", print.thres.best.method = "closest.topleft")

result.coords <- coords(results.roc,"best", best.method = "closest.topleft")
print(result.coords)

## threshold specificity sensitivity
## 1 0.5120248 0.6449093 0.7100291

datasettrain <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/train.set2.csv", show_col_types = FALSE)
datasettest <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/test.set2.csv", show_col_types = FALSE)
```r
train.set <- datasettrain
test.set <- datasettest

train.set <- train.set %>%
  mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))

test.set <- test.set %>%
  mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))

train.set = train.set %>% mutate_if(is.character, as.factor)
test.set = test.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 51342   49
dim(test.set)
## [1] 145016   49

nrow(train.set)
## [1] 51342
table(train.set$od_case)
##
##    No   Yes
## 25671 25671
dim(train.set)
## [1] 51342   49
dim(test.set)
## [1] 145016   49
table(test.set$od_case)
##
##    No   Yes
## 134008 11008
table(test.set$od_case)
##
##    No   Yes
## 134008 11008
class(train.set$od_case)
## [1] "factor"
```
```r
class(test.set$od_case)
## [1] "factor"

train.set <- train.set %>%
  select(od_case, ost, mood_anx_date_b, substancerelated.disorders, Stimulantuse, psychoticdisorders, Pol
ysubstance, Personalitydisorders, Opioiduse, Multiplementalillness, Mooddisorders, Intellectualdisability, Hallucinogensuse, earlyonsetdisorders, Developmentdisorders, Cocaineuse, Alcoholuse)

test.set <- test.set %>%
  select(od_case, ost, mood_anx_date_b, substancerelated.disorders, Stimulantuse, psychoticdisorders, Pol
ysubstance, Personalitydisorders, Opioiduse, Multiplementalillness, Mooddisorders, Intellectualdisability, Hallucinogensuse, earlyonsetdisorders, Developmentdisorders, Cocaineuse, Alcoholuse)

grid_default <- expand.grid(
  nrounds = 300,
  max_depth = 5,
  eta = 0.3, gamma = 0.001,
  colsample_bytree = 0.75,
  min_child_weight = 1,
  subsample = 1
)

train_control <- caret::trainControl(
  method = "none",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
  allowParallel = TRUE
)

model = caret::train(od_case ~ .,
  data = train.set,
  trControl = train_control,
  method = "xgbTree",
  tuneGrid = grid_default,
  verbose = TRUE
)

pred <- predict(model, test.set)
confusionMatrix(factor(pred), factor(test.set$od_case), positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference
## Prediction No Yes
##   No  115947 1815
##   Yes 18061 9193
```
##                Accuracy : 0.8629
##                  95% CI : (0.8612, 0.8647)
##     No Information Rate : 0.9241
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.4175
##
##  McNemar's Test P-Value : <2e-16
##
##     Sensitivity : 0.83512
##     Specificity : 0.86522
##     Pos Pred Value : 0.33731
##     Neg Pred Value : 0.98459
##     Prevalence : 0.07591
##     Detection Rate : 0.06339
##     Detection Prevalence : 0.18794
##     Balanced Accuracy : 0.85017
##
## 'Positive' Class : Yes
##
##
## prob <- predict(model, test.set, type = "prob")
## results.roc <- roc(test.set$od_case, prob$Yes)
## auc(results.roc)
##
## Area under the curve: 0.8922

plot(results.roc, print.thres = "best", print.thres.best.method = "closest.topleft")
results.coords <- coords(results.roc,"best", best.method = "closest.topleft")
print(results.coords)

## threshold specificity sensitivity
## 1 0.4906703   0.8651125   0.8354833

Validating all the models

datasettrain <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/train.set2.csv", show_col_types = FALSE)
datasettest <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/test.set2.csv", show_col_types = FALSE)

train.set <- datasettrain
test.set <- datasettest

train.set <- train.set %>%
  mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))
test.set <- test.set %>%
  mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))

train.set = train.set %>% mutate_if(is.character, as.factor)
test.set = test.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 51342  49
dim(test.set)
## [1] 145016  49

nrow(train.set)
## [1] 51342
table(train.set$od_case)
##
##   No  Yes
## 25671 25671

dim(train.set)
## [1] 51342  49
dim(test.set)
## [1] 145016  49
table(train.set$od_case)
```r
library(caret)

# XGBoost

grid_default <- expand.grid(
  nrounds = 300,
  max_depth = 5,
  eta = 0.3, gamma = 0.001,
  colsample_bytree = 0.75,
  min_child_weight = 1,
  subsample = 1
)

train_control <- caret::trainControl(
  method = "cv",
  number = 10,
  search = "grid",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
  allowParallel = TRUE
)

class(train.set$Sod_case)
## [1] "factor"

XGBoostmodel = caret::train(od_case ~ .,
  data = train.set,
  trControl = train_control,
  method = "xgbTree",
  tuneGrid = grid_default,
  verbose = TRUE
)
### Cross-Validated (10 fold) Confusion Matrix

(Entries are percentual average cell counts across resamples)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>44.6</td>
<td>8.9</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>5.4</td>
<td>41.1</td>
<td></td>
</tr>
</tbody>
</table>

**Accuracy (average): 0.8569**

### Confusion Matrix and Statistics

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>22897</td>
<td>4572</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>2774</td>
<td>21099</td>
<td></td>
</tr>
</tbody>
</table>

**Accuracy: 0.8569**

**95% CI: (0.8539, 0.8599)**

**No Information Rate: 0.5**

**P-Value [Acc > NIR]: < 2.2e-16**

**Kappa: 0.7138**

**Mcnemar's Test P-Value: < 2.2e-16**

**Sensitivity: 0.8919**

**Specificity: 0.8219**

**Pos Pred Value: 0.8336**

**Neg Pred Value: 0.8838**

**Prevalence: 0.5000**

**Detection Rate: 0.4460**

**Detection Prevalence: 0.5350**

**Balanced Accuracy: 0.8569**

'Positive' Class: No

---

### Random Forest

```r
grid_default <- expand.grid(.mtry = c(1:7), .splitrule = 'gini', .min.node.size = c(1))
```

```r
train_control <- caret::trainControl()
```
method = "cv",
number = 10,
search = "grid",
summaryFunction = twoClassSummary,
returnResamp = "none",
classProbs = TRUE,
savePredictions = TRUE,
verboseIter = FALSE,
allowParallel = TRUE
)

rf1 = caret::train(od_case ~ .,
data = train.set,
trControl = train_control,
method = "ranger",
tuneGrid = grid_default,
verbose = TRUE
)

print(rf1)
## Random Forest
##
## 51342 samples
##    48 predictor
##     2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 46208, 46206, 46208, 46208, 46208, 46208, ...
## Resampling results across tuning parameters:
##
##   mtry  ROC        Sens       Spec
##   1     0.9038600  0.9215068  0.7469122
##   2     0.9154990  0.9067039  0.8067470
##   3     0.9175635  0.8975107  0.8213938
##   4     0.9177619  0.8942774  0.8249776
##   5     0.9171787  0.8940827  0.8247439
##   6     0.9165390  0.8936931  0.8240428
##   7     0.9158401  0.8941606  0.8234974
##
## Tuning parameter 'splitrule' was held constant at a value of gini
## Tuning parameter 'min.node.size' was held constant at a value of 1
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were mtry = 4, splitrule = gini
## and min.node.size = 1.

classProbs = TRUE

confusionMatrix(rf1)
## Cross-Validated (10 fold) Confusion Matrix
##
### (entries are percentual average cell counts across resamples)
###
### Reference
### Prediction  No  Yes
###       No  44.7  8.8
###       Yes  5.3 41.2
###
### Accuracy (average) : 0.8596

confusionMatrix(rf1$pred, rf1$obs)

### Confusion Matrix and Statistics
###
### Reference
### Prediction     No    Yes
###       No  161777  34083
###       Yes  17920 145614
###
###                Accuracy : 0.8553
###                  95% CI : (0.8541, 0.8565)
###     No Information Rate : 0.5
###     P-Value [Acc > NIR] : < 2.2e-16
###
###                   Kappa : 0.7106
###
### McNemar's Test P-Value : < 2.2e-16
###
###     Sensitivity : 0.9003
###     Specificity : 0.8103
###   Pos Pred Value : 0.8260
###  Neg Pred Value : 0.8904
###    Prevalence : 0.5000
### Detection Rate : 0.4501
### Detection Prevalence : 0.5450
### Balanced Accuracy : 0.8553
###
### 'Positive' Class : No
###

#Adabag

grid_default <- expand.grid(
    mfinal = 300,
    maxdepth = 5
)

train_control <- caret::trainControl(
    method = "cv",
    number = 10,
    search = "grid",  
)
summaryFunction = twoClassSummary,
returnResamp = "none",
classProbs = TRUE,
savePredictions = TRUE,
verboseIter = FALSE,
allowParallel = TRUE
)

ada = caret::train(od_case ~ .,
data = train.set,
trControl = train_control,
method = "AdaBag",
tuneGrid = grid_default,
verbose = TRUE
)

print(ada)

## Bagged AdaBoost
##
##  51342 samples
##     48 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 46207, 46208, 46208, 46208, 46208, 46208, ...
## Resampling results:
##
##   ROC        Sens       Spec
##   0.8728535  0.8913562  0.820225
##
## Tuning parameter 'mfinal' was held constant at a value of 300
## Tuning parameter 'maxdepth' was held constant at a value of 5

confusionMatrix(ada)

## Cross-Validated (10 fold) Confusion Matrix
## (entries are percentual average cell counts across resamples)
##
##  Reference  No  Yes
## Prediction
##        No  44.6  9.0
##        Yes  5.4 41.0
##
## Accuracy (average) : 0.8558

cconfusionMatrix(ada$pred, ada$predsobs)
## Confusion Matrix and Statistics

<table>
<thead>
<tr>
<th>Prediction</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference</strong></td>
<td>22882</td>
<td>4615</td>
</tr>
<tr>
<td><strong>No</strong></td>
<td>2789</td>
<td>21056</td>
</tr>
</tbody>
</table>

---

### Accuracy: 0.8558
### 95% CI: (0.8527, 0.8588)
### No Information Rate: 0.5
### P-Value [Acc > NIR]: < 2.2e-16
### Kappa: 0.7116
### Mcnemar's Test P-Value: < 2.2e-16

### Sensitivity: 0.8914
### Specificity: 0.8202
### Pos Pred Value: 0.8322
### Neg Pred Value: 0.8830
### Prevalence: 0.5000
### Detection Rate: 0.4470
### Detection Prevalence: 0.5356
### Balanced Accuracy: 0.8558

'Positive' Class: No

---

```r
#SVM

train_control <- caret::trainControl(
    method = "cv",
    number = 10,
    search = "grid",
    summaryFunction = twoClassSummary,
    returnResamp = "none",
    classProbs = TRUE,
    savePredictions = TRUE,
    verboseIter = FALSE,
    allowParallel = TRUE
)

svmmodel = caret::train(od_case ~ .,
    data = train.set,
    trControl = train_control,
    method = "svmLinear",
    verbose = TRUE,
    tuneLength = 10
)

print(svmmodel)
```

```
## Support Vector Machines with Linear Kernel

### 51342 samples
### 48 predictor
### 2 classes: 'No', 'Yes'

### No pre-processing
### Resampling: Cross-Validated (10 fold)
### Summary of sample sizes: 46208, 46208, 46208, 46208, 46208, 46207, ...
### Resampling results:

### ROC  Sens  Spec
### 0.8934179  0.8713335  0.8279382

### Tuning parameter 'C' was held constant at a value of 1

```r
confusionMatrix(svmmodel)
```

### Cross-Validated (10 fold) Confusion Matrix
### (entries are percentual average cell counts across resamples)

```r
table <- confusionMatrix(svmmodel$pred, svmmodel$pred)
print(table)
```

```r
confusionMatrix(svmmodel$pred, svmmodel$obs)
```

### Confusion Matrix and Statistics

```r
table <- confusionMatrix(svmmodel$pred, svmmodel$obs)
print(table)
```

```
#           Reference
# Prediction  No  Yes
#        No 43.6  8.6
#        Yes 6.4 41.4

## Accuracy (average) : 0.8496
```
Detection Rate: 0.4357
Detection Prevalence: 0.5217
Balanced Accuracy: 0.8496

'Positive' Class: No

# Logistic Regression

```r
grid_default <- expand.grid(parameter = c(0.001, 0.01, 0.1, 1, 10, 100, 1000))

train_control <- caret::trainControl(
    method = "cv",
    number = 10,
    search = "grid",
    summaryFunction = twoClassSummary,
    returnResamp = "none",
    classProbs = TRUE,
    savePredictions = TRUE,
    verboseIter = FALSE,
    allowParallel = TRUE
)

LogitModel <- caret::train(od_case ~ .,
    data = train.set,
    method = "glm",
    family = "binomial",
    trControl = train_control,
    tuneGrid = grid_default
)
```

```
print(LogitModel)

## Generalized Linear Model
##
## 51342 samples
## 48 predictor
## 2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 46208, 46208, 46208, 46208, 46208, 46208, ...
## Resampling results across tuning parameters:
##
## ROC  Sens  Spec  parameter
## 0.9135732  0.8901485  0.821822  0.001
## 0.9135732  0.8901485  0.821822  0.001
## 0.9135732  0.8901485  0.821822  0.001
## 0.9135732  0.8901485  0.821822  0.001
## 0.9135732  0.8901485  0.821822  0.001
```
## ROC was used to select the optimal model using the largest value. The final value used for the model was parameter = 0.001.

`confusionMatrix(LogitModel)`

## Cross-Validated (10 fold) Confusion Matrix
## (entries are percentual average cell counts across resamples)
##
## Reference
## Prediction No Yes
## No 44.5 8.9
## Yes 5.5 41.1
##
## Accuracy (average) : 0.856

`confusionMatrix(LogitModel, LogitModel, LogitModel, LogitModel, LogitModel)`

## Confusion Matrix and Statistics
##
## Reference
## Prediction No Yes
## No 159957 32018
## Yes 19740 147679
##
## Accuracy : 0.856
## 95% CI : (0.8548, 0.8571)
## No Information Rate : 0.5
## P-Value [Acc > NIR] : < 2.2e-16
##
## Kappa : 0.712
##
## McNemar's Test P-Value : < 2.2e-16
##
## Sensitivity : 0.8901
## Specificity : 0.8218
## Pos Pred Value : 0.8332
## Neg Pred Value : 0.8821
## Prevalence : 0.5000
## Detection Rate : 0.4451
## Detection Prevalence : 0.5342
## Balanced Accuracy : 0.8560
##
## 'Positive' Class : No
##
## #GBM
train_control <- caret::trainControl(
  method = "cv",
  number = 10,
  search = "grid",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
  allowParallel = TRUE
)

gbm <- caret::train(od_case ~ .,
  data = train.set,
  method = "gbm",
  trControl = train_control,
)

print(gbm)

## Stochastic Gradient Boosting
##
## 51342 samples
##  48 predictor
##  2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 46208, 46208, 46208, 46207, 46207, 46208, ...
## Resampling results across tuning parameters:
##
##   interaction.depth  n.trees  ROC        Sens       Spec
##   1                   50      0.9024160  0.8789684  0.8137591
##   1                  100      0.9081720  0.8740601  0.8307042
##   1                  150      0.9120258  0.8784618  0.8294577
##   2                   50      0.9083011  0.8719565  0.8332362
##   2                  100      0.9137730  0.8801758  0.8330026
##   2                  150      0.9152219  0.8858243  0.8308991
##   3                   50      0.9123934  0.8757351  0.8363916
##   3                  100      0.9137730  0.8801758  0.8330026
##   3                  150      0.9167294  0.8887849  0.8303927
##
## Tuning parameter 'shrinkage' was held constant at a value of 0.1
##
## Tuning parameter 'n.minobsinnode' was held constant at a value of 10
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were n.trees = 150, interaction.depth =
## 3, shrinkage = 0.1 and n.minobsinnode = 10.

confusionMatrix(gbm)
### Cross-Validated (10 fold) Confusion Matrix
### (entries are percentual average cell counts across resamples)
###
### Reference
### Prediction  No  Yes
### No 44.4 8.5
### Yes 5.6 41.5
###
### Accuracy (average) : 0.8596

**confusionMatrix**(gbm$pred, gbm$obs)

### Confusion Matrix and Statistics
###
### Reference
### Prediction  No  Yes
### No 203263 39245
### Yes 27776 191794
###
### Accuracy : 0.855
### 95% CI : (0.8539, 0.856)
### No Information Rate : 0.5
### P-Value [Acc > NIR] : < 2.2e-16
###
### Kappa : 0.7099
###
### Mcnemar's Test P-Value : < 2.2e-16
###
### Sensitivity : 0.8798
### Specificity : 0.8301
### Pos Pred Value : 0.8382
### Neg Pred Value : 0.8735
### Prevalence : 0.5000
### Detection Rate : 0.4399
### Detection Prevalence : 0.5248
### Balanced Accuracy : 0.8550
###
### 'Positive' Class : No
###

#Naive Bayes

train_control <- caret::trainControl(
  method = "cv",
  number = 10,
  search = "grid",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
allowParallel = TRUE

nBmodel <- caret::train(od_case ~ ., 
  data = train.set, 
  method = "nb", 
  trControl = train_control
)

confusionMatrix(nBmodel)

## Cross-Validated (10 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
## Reference
## Prediction  No  Yes
##        No  49.8 39.9
##        Yes  0.2 10.1
##
## Accuracy (average) : 0.5995

confusionMatrix(nBmodel$pred, nBmodel$obs)

## Confusion Matrix and Statistics
##
## Reference
## Prediction  No  Yes
##        No  49034 27812
##        Yes  2308 23530
##
## Accuracy : 0.7067
## 95% CI : (0.7039, 0.7095)
## No Information Rate : 0.5
## P-Value [Acc > NIR] : < 2.2e-16
##
## Kappa : 0.4133
##
## McNemar's Test P-Value : < 2.2e-16
##
## Sensitivity : 0.9550
## Specificity : 0.4583
## Pos Pred Value : 0.6381
## Neg Pred Value : 0.9107
## Prevalence : 0.5000
## Detection Rate : 0.4775
## Detection Prevalence : 0.7484
## Balanced Accuracy : 0.7067
##
## 'Positive' Class : No
#KNN

grid_default <- expand.grid(k = 1)

train_control <- caret::trainControl(
  method = "cv",
  number = 10,
  search = "grid",
  summaryFunction = twoClassSummary,
  returnResamp = "none",
  classProbs = TRUE,
  savePredictions = TRUE,
  verboseIter = FALSE,
  allowParallel = TRUE
)

knnmodel <- caret::train(od_case ~ ,
  data = train.set,
  method = "knn",
  tuneGrid = grid_default,
  trControl = train_control)

confusionMatrix(knnmodel)

## Cross-Validated (10 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
## Reference
## Prediction  No  Yes
##  No  43.8 10.5
##  Yes  6.2 39.5
##
## Accuracy (average) : 0.8328

confusionMatrix(knnmodel$pred, knnmodel$obs)

## Confusion Matrix and Statistics
##
## Reference
## Prediction  No  Yes
##  No  22474 5387
##  Yes  3197 20284
##
## Accuracy : 0.8328
## 95% CI : (0.8296, 0.836)
## No Information Rate : 0.5
## P-Value [Acc > NIR] : < 2.2e-16
##
## Kappa : 0.6656
##
## McNemar's Test P-Value : < 2.2e-16
# Testing for the best performing model

```r
# Xgboost
predt <- predict(XGBoostmodel, test.set)
confusionMatrix(predt, test.set$od_case, positive = "Yes")
```

```r
## Confusion Matrix and Statistics
##
##                  Reference
## Prediction     No    Yes
##        No  119372   2016
##        Yes  14636   8992
##
##                Accuracy : 0.8852
##                 95% CI : (0.8835, 0.8868)
##        No Information Rate : 0.9241
##        P-Value [Acc > NIR] : 1
##
##                Kappa : 0.4637
##
## McNemar's Test P-Value : <2e-16
##
##                Sensitivity : 0.81686
##                Specificity : 0.89078
##                Pos Pred Value : 0.38057
##                Neg Pred Value : 0.98339
##                Prevalence : 0.07591
##                Detection Rate : 0.06201
##          Detection Prevalence : 0.16293
##           Balanced Accuracy : 0.85382
##
## 'Positive' Class : Yes
##
## prob <- predict(XGBoostmodel, test.set, type = "prob")
## results.roc <- roc(test.set$od_case, prob$Yes)
## auc(results.roc)
##
## Area under the curve: 0.9084
```
plot(results.roc, print.thres = "best", print.thres.best.method = "closest.topleft")

results.coords <- coords(results.roc,"best", best.method = "closest.topleft")
print(results.coords)

##   threshold specificity sensitivity
## 1 0.4371451   0.8774402   0.8272166

#RF
predrf <- predict(rf1, test.set)
confusionMatrix(factor(predrf), factor(test.set$sod_case), positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference
## Prediction No Yes
##   No     119745 2027
##   Yes    14263  8981
##
## Accuracy : 0.8877
## 95% CI : (0.886, 0.8893)
## No Information Rate : 0.9241
## P-Value [Acc > NIR] : 1
##
## Kappa : 0.4698
##
## McNemar's Test P-Value : <2e-16
##
## Sensitivity : 0.81586
## Specificity : 0.89357
## Pos Pred Value : 0.38638
## Neg Pred Value : 0.98335
## Prevalence : 0.07591
## Detection Rate : 0.06193
## Detection Prevalence : 0.16029
## Balanced Accuracy : 0.85471
##
## 'Positive' Class : Yes
##
prob <- predict(rf1, test.set, type = "prob")
results.roc1 <- roc(test.set$od_case, prob$Yes)
auc(results.roc1)

## Area under the curve: 0.9112

plot(results.roc1, print.thres = "best", print.thres.best.method = "closest.topleft")

#AdaBag
adat <- predict(ada, test.set)
confusionMatrix(adat, test.set$od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference
## Prediction | No | Yes |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>120482</td>
<td>2154</td>
</tr>
<tr>
<td>Yes</td>
<td>13526</td>
<td>8854</td>
</tr>
</tbody>
</table>

### Accuracy: 0.8919

95% CI: (0.8903, 0.8935)

### No Information Rate: 0.9241

### P-Value [Acc > NIR]: 1

### Kappa: 0.4772

### McNemar's Test P-Value: <2e-16

### Sensitivity: 0.80432

### Specificity: 0.89907

### Pos Pred Value: 0.39562

### Neg Pred Value: 0.98244

### Prevalence: 0.07591

### Detection Rate: 0.06106

### Detection Prevalence: 0.15433

### Balanced Accuracy: 0.85169

### 'Positive' Class: Yes

```r
prob <- predict(ada.test.set, type = "prob")
results.roc2 <- roc(test.set$Sod_case, prob$Yes)

auc(results.roc2)

## Area under the curve: 0.8686

plot(results.roc2, print.thres = "best", print.thres.best.method = "closest.topleft")
```
#Support Vector Machine

```
svmpt <- predict(svmmodel, test.set)
confusionMatrix(svmpt, test.set$od_case, positive = "Yes")
```

## Confusion Matrix and Statistics

```
##                   Reference
## Prediction     No    Yes
##        No  116205   1992
##        Yes  17803   9016
##
##                Accuracy : 0.8635
##                  95% CI : (0.8617, 0.8653)
## No Information Rate : 0.9241
## P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.4136
##
## Mcnemar's Test P-Value : <2e-16
##
## Sensitivity : 0.81904
## Specificity : 0.86715
## Pos Pred Value : 0.33618
## Neg Pred Value : 0.98315
## Prevalence : 0.07591
## Detection Rate : 0.06217
## Detection Prevalence : 0.18494
## Balanced Accuracy : 0.84310
```
## 'Positive' Class : Yes

```r
prob <- predict(svmmodel, test.set, type = "prob")
results.roc3 <- roc(test.set$od_case, prob$Yes)
auc(results.roc3)

## Area under the curve: 0.8805
```

```r
plot(results.roc3, print.thres = "best", print.thres.best.method = "closest.topleft")
```

---

# Logistic Regression

```r
LogitModelt <- predict(LogitModel, newdata = test.set)
confusionMatrix(LogitModelt, test.set$od_case, positive = "Yes")
```

## Confusion Matrix and Statistics

```
## Confusion Matrix and Statistics
##                Reference
## Prediction     No  Yes
##        No 118866  2037
##        Yes  15142  8971
##
##                Accuracy : 0.8815
##                  95% CI : (0.8799, 0.8832)
##                No Information Rate : 0.9241
##                P-Value [Acc > NIR] : 1
```
## Kappa : 0.4539
#
## Mcnemar's Test P-Value : <2e-16
##
## Sensitivity : 0.81495
## Specificity : 0.88701
## Pos Pred Value : 0.37204
## Neg Pred Value : 0.98315
## Prevalence : 0.07591
## Detection Rate : 0.06186
## Detection Prevalence : 0.16628
## Balanced Accuracy : 0.85098
#
## 'Positive' Class : Yes
##
prob <- \texttt{predict}(\texttt{LogitModel}, \texttt{test.set, type = "prob"})
results.roc4 <- \texttt{roc}(\texttt{test.set\$od_case, prob\$Yes})
auc(results.roc4)
## Area under the curve: 0.9067

\texttt{plot}(results.roc4, print.thres = "best", print.thres.best.method = "closest.topleft")

#GBM
gbmt <- \texttt{predict}(\texttt{gbm, newdata = test.set})
\texttt{confusionMatrix}(gbmt, \texttt{test.set\$od_case, positive = "Yes"})
## Confusion Matrix and Statistics

<table>
<thead>
<tr>
<th>Reference</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction</td>
<td>No</td>
<td>118853</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>15155</td>
</tr>
</tbody>
</table>

### Accuracy : 0.8821
### 95% CI : (0.8804, 0.8837)
### No Information Rate : 0.9241
### P-Value [Acc > NIR] : 1
### Kappa : 0.4579

### Mcnemar's Test P-Value : <2e-16

### Sensitivity : 0.82322
### Specificity : 0.88691
### Pos Pred Value : 0.37420
### Neg Pred Value : 0.98389
### Prevalence : 0.07591
### Detection Rate : 0.06249
### Detection Prevalence : 0.16700
### Balanced Accuracy : 0.85506

'Positive' Class : Yes

```
prob <- predict(gbm, test.set, type = "prob")
results.roc5 <- roc(test.set$od_case, prob$Yes)
auc(results.roc5)

# Area under the curve: 0.91
```

```
plot(results.roc5, print.thres = "best", print.thres.best.method = "closest.topleft")
```
#Naive Bayes

Nbpt <- predict(nBmodel, newdata = test.set)

confusionMatrix(Nbpt, reference = test.set$od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference  
## Prediction  No Yes  
##        No 133523 8786  
##        Yes  485 2222  
##
##                Accuracy : 0.9361  
##                  95% CI : (0.9348, 0.9373)  
##     No Information Rate : 0.9241  
##     P-Value [Acc > NIR] : < 2.2e-16  
##
##                   Kappa : 0.3031  
##
##  McNemar's Test P-Value : < 2.2e-16
##
##     Sensitivity : 0.20185  
##     Specificity : 0.99638  
##     Pos Pred Value : 0.82083  
##     Neg Pred Value : 0.93826  
##     Prevalence : 0.07591  
##     Detection Rate : 0.01867  
##     Detection Prevalence : 0.01867  

0.466 (0.878, 0.830)
## Balanced Accuracy : 0.59912
## 'Positive' Class : Yes
##
prob <- predict(nBmodel, test.set, type = "prob")
results.roc6 <- roc(test.set$sod_case, prob$Yes)
auc(results.roc6)
## Area under the curve: 0.8843

plot(results.roc6, print.thres = "best", print.thres.best.method = "closest.topleft")

##KNN
Knnpt <- predict(knnmodel, newdata = test.set)
confusionMatrix(Knnpt, test.set$sod_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference
##
##         Prediction No  Yes
##       No 117573 2384
##       Yes 16435 8624

##
## Accuracy : 0.8702
## 95% CI : (0.8685, 0.872)
## No Information Rate : 0.9241
## P-Value [Acc > NIR] : 1
### Kappa : 0.4167

### Mcnemar's Test P-Value : <2e-16

### Sensitivity : 0.78343
### Specificity : 0.87736
### Pos Pred Value : 0.34415
### Neg Pred Value : 0.98013
### Prevalence : 0.07591
### Detection Rate : 0.05947
### Detection Prevalence : 0.17280
### Balanced Accuracy : 0.83039

### 'Positive' Class : Yes

```r
prob <- predict(knnmodel, test.set, type = "prob")
results.roc7 <- roc(test.set$od_case, prob$Yes)
auc(results.roc7)

## Area under the curve: 0.8656

plot(results.roc7, print.thres = "best", printe.thres.best.method = "closest.topleft")
```

ML deep learning with Keras
library(keras)
library(tensorflow)
library(rlang)
library(tidyverse)
library(dplyr)
library(gapminder)
library(caret)
library(xgboost)
library(forcats)
library(mice)
library(vctrs)
library(naniar)
library(yardstick)
library(forcats)

datasettrain <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/train.set2.csv", show_col_types = FALSE)
datasettest <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/test.set2.csv", show_col_types = FALSE)

train.set <- datasettrain
test.set <- datasettest

train.set = train.set %>% mutate_if(is.character, as.factor)
test.set = test.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 51342  49
dim(test.set)
## [1] 145016  49

nrow(train.set)
## [1] 51342
table(train.set$od_case)
##    0   1
## 25671 25671
dim(train.set)
## [1] 51342  49
dim(test.set)
## [1] 145016  49
```r
table(train.set$od_case)
##
##   0  1
## 25671 25671

table(test.set$od_case)
##
##   0  1
## 134008 11008

class(train.set$od_case)
## [1] "numeric"
class(test.set$od_case)
## [1] "numeric"

x_train_tbl <- train.set %>%
    select(-od_case)
x_test_tbl <- test.set %>%
    select(-od_case)
y_train_vec <- train.set %>%
    select(od_case)
y_test_vec <- test.set %>%
    select(od_case)
glimpse(x_train_tbl)
## Rows: 51,342
## Columns: 48
## $ ost                                  <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ~
## $ rheumatoid_arthritis_date_b          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ parkinsonism_date_b                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ osteo_arthritis_date_b               <dbl> 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, ~
## $ osteo_porosis_date_b                 <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ~
## $ ms_date_b                            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ mood_anx_date_b                      <dbl> 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, ~
## $ isch_stroke_date_b                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ ihd_date_b                           <dbl> 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, ~
## $ hypertension_date_b                  <dbl> 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, ~
## $ hosp_tia_date_b                      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ hosp_stroke_date_b                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ heart_failure_date_b                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ~
## $ haemorr_stroke_date_b                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ epilepsy_date_b                      <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ~
## $ diabetes_date_b                      <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, ~
## $ depression_date_b                    <dbl> 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, ~
```
## $ copd_date_b  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ ckd_date_b   <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ~
## $ asthma_date_b <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, ~
## $ angina_date_b <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ ami_date_b    <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ~
## $ alzheimer_dementia_date_b <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ~
## $ Tobaccouse    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Tissueinfection <dbl> 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, ~
## $ substancerelated.disorders <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Stimulantuse <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Sepsis        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Personalitydisorders <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Otherpsychoactivedruguse <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Osteomyelitis <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Opioiduse <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Neuroticrelateddisorders <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Neurocognitivedisorders <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Multiplementalillness <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, ~
## $ Mooddisorders <dbl> 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, ~
## $ Intellectualdisability <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Hallucinogensuse <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Endocarditis <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ earlyonsetdisorders <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Developmentdisorders <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Cocaineuse <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Cannabinoiduse <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Behaviouralpsychologicaldisturbances <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Alcoholuse   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~

model_keras <- keras_model_sequential()

model_keras %>%
layer_dense(
  units = 49,
  activation = "relu",
  input_shape = ncol(x_train_tbl)) %>%
layer_dropout(rate = 0.001) %>%
layer_dense(
  units = 49,
  activation = "relu") %>%
layer_dropout(rate = 0.001) %>%
layer_dense(
  units = 49,
  activation = "relu") %>%
layer_dropout(rate = 0.001) %>%
layer_dense(
  units = 49,
  activation = "relu") %>%
layer_dropout(rate = 0.001) %>%
layer_dense(
  units = 49,
  activation = "relu") %>%
layer_dropout(rate = 0.001) %>%
layer_dense(
  units = 1,
  activation = "sigmoid") %>%
compile(
  optimizer = 'adam',
  loss = 'binary_crossentropy',
  metrics = c('accuracy'))
)

model_keras

## Model

### Layer (type)                        Output Shape                    Param #

### -----------------------------
### dense_5 (Dense)                 (None, 49)                      2401
### dropout_4 (Dropout)             (None, 49)                      0
### dense_4 (Dense)                 (None, 49)                      2450
### dropout_3 (Dropout)             (None, 49)                      0
### dense_3 (Dense)                 (None, 49)                      2450
### dropout_2 (Dropout)             (None, 49)                      0
### dense_2 (Dense)                 (None, 49)                      2450
### dropout_1 (Dropout)             (None, 49)                      0
### dense_1 (Dense)                 (None, 49)                      2450
### dropout (Dropout)               (None, 49)                      0
### dense (Dense)                   (None, 1)                       50

### Total params: 12,251
### Trainable params: 12,251
### Non-trainable params: 0

history <- fit(
  object = model_keras,
x = as.matrix(x_train_tbl),
y = as.matrix(y_train_vec),
batch_size = 300,
epochs = 1000,
validation_split = 0.3)

print(history)

### Trained on 35,939 samples (batch_size=300, epochs=1,000)
### Final epoch (plot to see history):
### val_loss: 1.309
### val_acc: 0.7212
### loss: 0.1665
### acc: 0.9342

plot(history)

yhat_keras_class_vec <- predict_classes(object = model_keras, x = as.matrix(x_test_tbl)) %>%
  as.vector()

yhat_keras_prob_vec <- predict_proba(object = model_keras, x = as.matrix(x_test_tbl)) %>%
  as.vector()

estimates_tbl <- y_test_vec %>%
  add_column(estimate = as.factor(yhat_keras_class_vec),
  class_prob = yhat_keras_prob_vec)
estimates_tbl

### A tibble: 145,016 x 3
### od_case estimate class_prob
### <dbl> <fct>         <dbl>
## 1       1 1         0.815
## 2       1 0         0.169
## 3       1 0         0.137
## 4       1 1         0.966
## 5       1 1         1.00
## 6       1 1         0.929
## 7       1 1         1.00
## 8       1 0         0.0000450
## 9       1 1         1
## 10       1 1         1
## # i 145,006 more rows

estimates_tbl = estimates_tbl %>% mutate_at(vars(od_case), list(factor))
class(estimates_tbl$od_case)
## [1] "factor"
levels(estimates_tbl$od_case)
## [1] "0" "1"
levels(estimates_tbl$estimate)
## [1] "0" "1"
options(yardstick.event_first = FALSE)
estimates_tbl %>% conf_mat(od_case, estimate)

### Truth
### Prediction 0 1
### 0 121695 3165
### 1 12313 7843

estimates_tbl %>% metrics(od_case, estimate)

### # A tibble: 2 x 3
### .metric .estimator .estimate
### <chr>   <chr>          <dbl>
### accuracy binary        0.893
### kap      binary         0.449

estimates_tbl %>% roc_auc(od_case, class_prob)

### # A tibble: 1 x 3
### .metric .estimator .estimate
ML modeling with Tidy models

```r
library(keras)
library(tensorflow)
library(rlang)
library(tidyverse)
library(dplyr)
library(gapminder)
library(xgboost)
library(forcats)
library(mice)
library(vctrs)
library(naniar)
library(yardstick)
library(forcats)
library(Boruta)
library(ROSE)
library(imbalance)
library(rnn)
library(ROCR)
library(SHAPforxgboost)
library(randomForest)
library(mlbench)
library(ggplot2)
library(data.table)
library(here)
library(tidymodels)
library(parsnip)
library(caret)
library(kknn)
library(yardstick)
library(tune)
```

### Testing the original sample

```r
dataset_train <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/train.set2.csv", show_col_types = FALSE)

dataset_test <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/test.set2.csv", show_col_types = FALSE)

train.set <- dataset_train
test.set <- dataset_test

train.set <- train.set %>%
  mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))

test.set <- test.set %>%
```
```r
mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))

train.set = train.set %>% mutate_if(is.character, as.factor)
test.set = test.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 51342  49
dim(test.set)
## [1] 145016  49

nrow(train.set)
## [1] 51342
table(train.set$od_case)
##
## No Yes
## 25671 25671
dim(train.set)
## [1] 51342  49
dim(test.set)
## [1] 145016  49
table(test.set$od_case)
##
## No Yes
## 134008 11008

class(train.set$od_case)
## [1] "factor"
class(test.set$od_case)
## [1] "factor"

#XGBoost

xgb_spec <- boost_tree(trees = 300) %>%
```
```r
set_engine("xgboost") %>%
set_mode("classification")

overdose_rec <-
recipe(od_case ~ ., data = train.set)

xgbworkflow <- workflow() %>%
  add_model(xgb_spec) %>%
  add_recipe(overdose_rec)

xgb_fit <- xgbworkflow %>%
  fit(data = train.set)

class_pred <- predict(xgb_fit, test.set)
prob_pred <- predict(xgb_fit, test.set, type = "prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("XGB_Class", "LR_DeathProb", "R_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)
caret::confusionMatrix(ODD_preds$XGB_Class, ODD_preds$od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference
## Prediction  No  Yes
##        No 119123  2035
##        Yes  14885  8973
##
## Accuracy : 0.8833
## 95% CI : (0.8817, 0.885)
## No Information Rate : 0.9241
## P-Value [Acc > NIR] : 1
##
## Kappa : 0.4585
##
## Mcnemar's Test P-Value : <2e-16
##
## Sensitivity : 0.81513
## Specificity : 0.88892
## Pos Pred Value : 0.37610
## Neg Pred Value : 0.98320
## Prevalence : 0.07591
## Detection Rate : 0.06188
## Detection Prevalence : 0.16452
## Balanced Accuracy : 0.85203
##
## 'Positive' Class : Yes
```

# Random Forest

rf_mod <-
  rand_forest(trees=500) %>%
  set_engine("ranger") %>%
  set_mode("classification")

rf_fit <-
  rf_mod %>%
  fit(od_case ~ ., data = train.set)

class_pred <- predict(rf_fit, test.set)
prob_pred <- predict(rf_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("RF_Class", "RF_DeathProb", "RF_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)

table(ODD_preds$od_case)

##
##     No    Yes
## 134008  11008

caret::confusionMatrix(ODD_preds$RF_Class, ODD_preds$od_case, positive = "Yes")

## Confusion Matrix and Statistics
##
## Reference  Prediction
##      No    119667   1994
##      Yes    14341   9014
##
##                Accuracy : 0.8874
##                  95% CI : (0.8857, 0.889)
##     No Information Rate : 0.9241
##     P-Value [Acc > NIR] : 1
##
##                Kappa : 0.4699
##
##  McNemar's Test P-Value : <2e-16
##
##                Sensitivity : 0.81886
##                Specificity : 0.89298
##                Pos Pred Value : 0.38596
##                Neg Pred Value : 0.98361
##                Prevalence : 0.07591
##                Detection Rate : 0.06216
##                Detection Prevalence : 0.16105
##                Balanced Accuracy : 0.85592
## 'Positive' Class : Yes

### feature selected RF (From Boruta)

```r
datasettrain <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/train.set2.csv", show_col_types = FALSE)
datasettest <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/test.set2.csv", show_col_types = FALSE)

train.set <- datasettrain
test.set <- datasettest

train.set <- train.set %>%
  mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))

test.set <- test.set %>%
  mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))

train.set = train.set %>% mutate_if(is.character, as.factor)
test.set = test.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 51342   49
dim(test.set)
## [1] 145016   49
nrow(train.set)
## [1] 51342
table(train.set$od_case)
##
##   No  Yes
## 25671 25671
dim(train.set)
## [1] 51342   49
dim(test.set)
## [1] 145016   49
table(train.set$od_case)
##
##   No  Yes
## 25671 25671
```
```r
table(test.set$od_case)
##
##    No   Yes
## 134008 11008

class(train.set$od_case)
## [1] "factor"

class(test.set$od_case)
## [1] "factor"

train.set <- train.set %>% select(od_case, alzheimer_dementia_date_b, ami_date_b, angina_date_b, ckd_date_b, copd_date_b, depression_date_b, diabetes_date_b, heart_failure_date_b, hypertension_date_b, hosp_stroke_date_b, haemorrhage_stroke_date_b, isch_stroke_date_b, ihd_date_b, mood_anxiety_date_b, osteo_arthritis_date_b, osteo_porosis_date_b, ost, substance_related_disorders, Tobacco_use, Cocaine_use, early_onset_disorders, Mood_disorders, Multiple_specialty_illness, Neurocognitive_disorders, Neurotic_related_disorders, Opioid_use, Osteomyelitis, Other_psychotropic_drug_use, Personality_disorders, Polysubstance, Psychotic_disorders, Sepsis, Stimulant_use, Titanium_infection)

test.set <- test.set %>% select(od_case, alzheimer_dementia_date_b, ami_date_b, angina_date_b, ckd_date_b, copd_date_b, depression_date_b, diabetes_date_b, heart_failure_date_b, hypertension_date_b, hosp_stroke_date_b, haemorrhage_stroke_date_b, isch_stroke_date_b, ihd_date_b, mood_anxiety_date_b, osteo_arthritis_date_b, osteo_porosis_date_b, ost, substance_related_disorders, Tobacco_use, Cocaine_use, early_onset_disorders, Mood_disorders, Multiple_specialty_illness, Neurocognitive_disorders, Neurotic_related_disorders, Opioid_use, Osteomyelitis, Other_psychotropic_drug_use, Personality_disorders, Polysubstance, Psychotic_disorders, Sepsis, Stimulant_use, Titanium_infection)

rf_mod <-
  rand_forest(trees=500) %>%
  set_engine("ranger") %>%
  set_mode("classification")

rf_fit <-
  rf_mod %>%
  fit(od_case ~., data = train.set)

class_pred <- predict(rf_fit, test.set)
prob_pred <- predict(rf_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("RF_Class", "RF_DeathProb", "RF_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)

caret::confusionMatrix(ODD_preds$RF_Class, ODD_preds$od_case, positive = "Yes")
## Confusion Matrix and Statistics
##
```
Validating all the models

datasettrain <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/train.set2.csv", show_col_types = FALSE)
datasettest <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/test.set2.csv", show_col_types = FALSE)

train.set <- datasettrain
test.set <- datasettest

train.set <- train.set %>%
  mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))
test.set <- test.set %>%
  mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))

train.set = train.set %>% mutate_if(is.character, as.factor)
test.set = test.set %>% mutate_if(is.character, as.factor)

dim(train.set)
## [1] 51342   49

dim(test.set)
```
## [1] 145016  49
nrow(train.set)
## [1] 51342
table(train.set$od_case)
##   No Yes
## 25671 25671
dim(train.set)
## [1] 51342  49
dim(test.set)
## [1] 145016  49
table(test.set$od_case)
##   No Yes
## 134008 11008
class(train.set$od_case)
## [1] "factor"
class(test.set$od_case)
## [1] "factor"

# Xgboost

ten_fold <- vfold_cv(train.set, v=10)
xgb_spec <- boost_tree(  
trees = 300)  
set_engine("xgboost")  
set_model("classification")

overdose_rec <-  
  recipe(od_case ~., data = train.set)

xgbworkflow <- workflow(  
  add_model(xgb_spec)
)  
```
add_recipe(overdose_rec)

xgb_fit <- xgbworkflow %>%
  fit(data = train.set)

xgb_fit_rs <-
  xgbworkflow %>%
  fit_resamples(ten_fold)

collect_metrics(xgb_fit_rs)

### # A tibble: 2 x 5
### .metric .estimator  mean     n std_err
### <chr>    <chr>      <dbl> <int>   <dbl>
### 1 accuracy binary     0.855    10 0.00103
### 2 roc_auc  binary     0.911    10 0.00113

#Random Forest

ten_fold <- vfold_cv(train.set, v=10)

rf_spec <- rand_forest(
  trees = 500) %>%
  set_engine("ranger") %>%
  set_mode("classification")

overdose_rec <-
  recipe(od_case ~., data = train.set)

rfworkflow <- workflow() %>%
  add_model(rf_spec) %>%
  add_recipe(overdose_rec)

rf_fit <- rfworkflow %>%
  fit(data = train.set)

rf_fit_rs <-
  rfworkflow %>%
  fit_resamples(ten_fold)

collect_metrics(rf_fit_rs)

### # A tibble: 2 x 5
### .metric .estimator  mean     n std_err
### <chr>    <chr>      <dbl> <int>   <dbl>
### 1 accuracy binary     0.859    10 0.00168
### 2 roc_auc  binary     0.917    10 0.00170

#Logistic Regression

ten_fold <- vfold_cv(train.set, v=10)
glm_spec <- logistic_reg() %>%
set_engine("glm") %>%
translate()

overdose_rec <-
  recipe(od_case ~., data = train.set)

glmworkflow <- workflow() %>%
  add_model(glm_spec) %>%
  add_recipe(overdose_rec)

glm_fit <- glmworkflow %>%
  fit(data = train.set)

glm_fit_rs <-
  glmworkflow %>%
  fit_resamples(ten_fold)

collect_metrics(glm_fit_rs)

## # A tibble: 2 x 5
## .metric  .estimator  mean     n std_err
## <chr>    <chr>      <dbl> <int>   <dbl>
## 1 accuracy binary     0.856    10 0.00168
## 2 roc_auc  binary     0.914    10 0.00127

#KNN
ten_fold <- vfold_cv(train.set, v=10)

knn_spec <- nearest_neighbor() %>%
  set_engine("kknn") %>%
  set_mode("classification") %>%
  translate

overdose_rec <-
  recipe(od_case ~., data = train.set)

knnworkflow <- workflow() %>%
  add_model(knn_spec) %>%
  add_recipe(overdose_rec)

knn_fit <- knnworkflow %>%
  fit(data = train.set)

collect_metrics(knn_fit_rs)
# A tibble: 2 x 5
## # .metric .estimator  mean  n  std_err
##  <chr>   <chr>      <dbl> <int>   <dbl>
## 1 accuracy binary     0.761    10 0.00139
## 2 roc_auc  binary     0.836    10 0.00203

# Testing for the best performing model

### XGBoost

class_pred <- predict(xgb_fit, test.set)
prob_pred <- predict(xgb_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("XGB_Class", "LR_DeathProb", "R_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)

cmxgboost <- caret::confusionMatrix(ODD_preds$XGB_Class, ODD_preds$od_case, positive = "Yes")

print(cmxgboost)

## Confusion Matrix and Statistics
## Reference
## Prediction  No  Yes
##        No 119123  2035
##        Yes 14885  8973
##
##                Accuracy : 0.8833
##                  95% CI : (0.8817, 0.885)
## No Information Rate : 0.9241
## P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.4585
##
## Mcnemar's Test P-Value : <2e-16
##
##   Sensitivity : 0.81513
## Specificity : 0.88892
## Pos Pred Value : 0.37610
## Neg Pred Value : 0.98320
## Prevalence : 0.07591
## Detection Rate : 0.06188
## Detection Prevalence : 0.16452
## Balanced Accuracy : 0.85203
##
## 'Positive' Class : Yes

### RF
class_pred <- predict(rf_fit, test.set)
prob_pred <- predict(rf_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("rf_Class", "LR_DeathProb", "R_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)

cmrf <- caret::confusionMatrix(ODD_preds$rfsrf_Class, ODD_preds$sod_case, positive = "Yes")
print(cmrf)
## Confusion Matrix and Statistics
## Reference
## Prediction No Yes
## No 119655 1990
## Yes 14353 9018
## Accuracy : 0.8873
## 95% CI : (0.8857, 0.8889)
## No Information Rate : 0.9241
## P-Value [Acc > NIR] : 1
## Kappa : 0.4699
## McNemar's Test P-Value : <2e-16
## Sensitivity : 0.81922
## Specificity : 0.89289
## Pos Pred Value : 0.38586
## Neg Pred Value : 0.98364
## Prevalence : 0.07591
## Detection Rate : 0.06219
## Detection Prevalence : 0.16116
## Balanced Accuracy : 0.85606
## 'Positive' Class : Yes
##
### GLM
class_pred <- predict(glm_fit, test.set)
prob_pred <- predict(glm_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("glm_Class", "LR_DeathProb", "R_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)
cm <- caret::confusionMatrix(ODD_preds$glm_Class, ODD_preds$od_case, positive = "Yes")

print(cm)

## Confusion Matrix and Statistics
## Reference
## Prediction  No  Yes
##        No 118866 2037
##        Yes 15142  8971
##
##                Accuracy : 0.8815
##                  95% CI : (0.8799, 0.8832)
##     No Information Rate : 0.9241
##     P-Value [Acc > NIR] : 1
##
##                Kappa : 0.4539
##
## McNemar's Test P-Value : <2e-16
##
##                Sensitivity : 0.81495
##                Specificity : 0.88701
##             Pos Pred Value : 0.37204
##             Neg Pred Value : 0.98315
##              Prevalence : 0.07591
##           Detection Rate : 0.06186
##    Detection Prevalence : 0.16628
##        Balanced Accuracy : 0.85098
##
## 'Positive' Class : Yes
##
##### KNN

class_pred <- predict(knn_fit, test.set)
prob_pred <- predict(knn_fit, test.set, type="prob")
predictions <- data.frame(class_pred, prob_pred) %>%
  setNames(c("knn_Class", "LR_DeathProb", "R_NotDeathProb"))

ODD_preds <- test.set %>%
  bind_cols(predictions)

knn_cm <- caret::confusionMatrix(ODD_preds$knn_Class, ODD_preds$od_case, positive = "Yes")

print(knn_cm)

## Confusion Matrix and Statistics
## Reference
## Prediction

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No</strong></td>
<td>121556</td>
<td>4332</td>
</tr>
<tr>
<td><strong>Yes</strong></td>
<td>12452</td>
<td>6676</td>
</tr>
</tbody>
</table>

## Accuracy

- Accuracy: 0.8843
- 95% CI: (0.8826, 0.8859)

## No Information Rate

- No Information Rate: 0.9241

## P-Value [Acc > NIR]

- P-Value: 1

## Kappa

- Kappa: 0.3837

## McNemar's Test P-Value

- McNemar's Test P-Value: <2e-16

## Sensitivity

- Sensitivity: 0.60647

## Specificity

- Specificity: 0.90708

## Pos Pred Value

- Pos Pred Value: 0.34902

## Neg Pred Value

- Neg Pred Value: 0.96559

## Prevalence

- Prevalence: 0.07591

## Detection Rate

- Detection Rate: 0.04604

## Detection Prevalence

- Detection Prevalence: 0.13190

## Balanced Accuracy

- Balanced Accuracy: 0.75677

## 'Positive' Class

- 'Positive' Class: Yes

---

Inference code

```r
### Shap XGBoost now

library(data.table)
library(tidyverse)
library(caret)
library(xgboost)
library(SHAPforxgboost)
library(here)
library(ggplot2)
library(Boruta)
library(randomForest)
library(naniar)

fullset <- read_csv("U:/Andy's Thesis/MLoverdose in general/GeneralodML/inferencedataset1.csv", show_col_types = FALSE)

# Check the distribution of the categorical variable

table(fullset$od_case)

# Count of 0 and 1

## 0   1
## 446710 36679

fullset$od_case <- as.factor(fullset$od_case)

class(fullset$od_case)
```

521
Logit Regression

glmmodel <- glm(od_case ~ ., data = fullset, family = binomial(link = "logit"))

summary(glmmodel)

## Call:
## glm(formula = od_case ~ ., family = binomial(link = "logit"),
##     data = fullset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8705 -0.2146  -0.1883  -0.1381   3.4204

## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -1.840185  0.044511 -41.342  < 2e-16 ***
## ost                       0.930037  0.024851  37.425  < 2e-16 ***
## rheumatoid_arthritis_date_b -0.104036  0.050600  -2.056  0.03978
## parkinsonism_date_b       -0.084999  0.112333  -0.757  0.44925
## osteo_arthritis_date_b    -0.182199  0.021371  -8.526  < 2e-16 ***
## osteo_porosis_date_b     -0.341159  0.039634  -8.608  < 2e-16 ***
## ms_date_b                 -0.144621  0.113038  -1.279  0.20075
## mood_anx_date_b          -0.033289  0.027497  -1.211  0.22604
## isch_stroke_date_b       -0.499083  0.285333  -1.749  0.08027
## ihd_date_b               -0.004185  0.033794  -0.124  0.90143
## hypertension_date_b      -0.542089  0.023945 -22.639  < 2e-16 ***
## hosp_tia_date_b          0.039683  0.096030   0.413  0.67943
## hosp_stroke_date_b       0.890121  0.290753   3.061  0.00220
## heart_failure_date_b     0.154473  0.037062   4.168  3.07e-05 ***
## haemorr_stroke_date_b    0.096377  0.275477   0.350  0.72645
## epilepsy_date_b          0.257154  0.039574   6.498  3.57e-11 ***
## diabetes_date_b          -0.199710  0.023826  -8.382  < 2e-16 ***
## depression_date_b        -0.049145  0.026415  -1.861  0.06281
## copd_date_b               0.092109  0.026998   3.412  0.00064
## ckd_date_b                0.216011  0.029167   7.405  1.31e-13 ***
## asthma_date_b            0.100549  0.037066   2.720  0.00647
## angina_date_b            -0.054784  0.048346  -1.133  0.25715
## ami_date_b                0.113245  0.050660   2.235  0.02539
## alzheimer_dementia_date_b 0.565452  0.050564  -1.183  0.23904
## Tobaccouse                -0.706672  0.031320  -22.563  < 2e-16 ***
## Tissueinfection           -0.292123  0.019508  -14.975  < 2e-16 ***
## substancerelated.disorders 2.360938  0.018017  131.037  < 2e-16 ***
## Stimulantuse              0.636755  0.033429   19.048  < 2e-16 ***
## Sepsis                    0.482848  0.025454   18.969  < 2e-16 ***
## Sedativeandhypnoticuse    0.023039  0.057066   0.404  0.68640
## psychoticdisorders        0.123765  0.025868   4.784  1.71e-06 ***
## Polysubstance             0.892725  0.021832  40.890  < 2e-16 ***
## Personality disorders                  0.301841   0.025715  11.738  < 2e-16 ***
## Other psychoactive drug use              0.292309   0.031527   9.272  < 2e-16 ***
## Osteomyelitis                           -0.063105   0.040797  -1.547 0.121911
## Opioid use                              1.392091   0.021629  64.361  < 2e-16 ***
## Neurotic related disorders              -0.145990   0.016499  -8.848  < 2e-16 ***
## Neurocognitive disorders                0.104382   0.025423   4.106 4.03e-05 ***
## Multiplemental illness                  -0.092108   0.016679  -5.523 3.34e-08 ***
## Mood disorders                          -0.006004   0.018862  -0.318 0.750257
## Intellectual disability                -0.099731   0.096450  -1.034 0.301131
## Hallucinogens use                       0.458135   0.127729   3.587 0.000335 ***
## Endocarditis                            -0.082062   0.050299  -1.632 0.102785
## Early onset disorders                   0.062195   0.020792   2.991 0.002778 **
## Development disorders                   0.063933   0.011101   5.523 0.210968
## Cocaine use                             0.562113   0.028829  19.498  < 2e-16 ***
## Cannabinoids use                        -0.292788   0.031157  -9.397  < 2e-16 ***
## Behavioural psychological disturbances -0.089811   0.074360  -1.208 0.227128
## Alcohol use                             -1.845672   0.045373  -40.678  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Dispersion parameter for binomial family taken to be 1)
## Null deviance: 259664  on 483388 degrees of freedom
## Residual deviance: 148007  on 483340 degrees of freedom
## AIC: 148105
## Number of Fisher Scoring iterations: 7

plot(glmmodel)
# oddsratio <- exp(cbind(Odds_Ratio = coef(glmmodel), confint(glmmodel)))
#
# oddsratio
oddsratio <- read_csv("U:/Andy's Thesis/MLOverdose in general/GeneralodML/oddsratiood.csv", show_col_types = FALSE)

oddsratio

## # A tibble: 49 x 4
## # Variables                         Odds_Ratio SDE2.5 SDE975
## <chr>                        <dbl>  <dbl>  <dbl>
## 1 Alcohol_use                     0.158  0.145  0.173
## 2 Alzheimer_Dementia                0.568  0.514  0.627
## 3 Acute_Myocardial_Infraction       1.12   1.01   1.24
## 4 Angina                                0.947  0.861  1.04
## 5 Asthma                                1.11   1.07   1.14
## 6 Behavioural_psychological_disturbances 0.914  0.790  1.06
## 7 Cannabinoid_use                      0.746  0.702  0.793
## 8 Chronic_Kidney_Disease             1.24   1.17   1.31
## 9 Cocaine_use                         1.75   1.66   1.86
## 10 Chronic Obstructive Pulmonary Disease 1.10   1.04   1.16
## 11 Depression                          0.952  0.904  1.00
## 12 Developmentdisorders               1.07   0.964  1.18
## 13 Diabetes                             0.819  0.782  0.858
## 14 Earlyonset_disorders                1.06   1.02   1.11
## 15 Endocarditis                        0.921  0.834  1.02
## 16 Epilepsy                            1.29   1.20   1.40
## 17 Haemorrhagic_Stroke                 1.10   0.628  1.85
## 18 Hallucinogensuse                     1.58   1.23   2.04
## 19 Heart_Failure                       1.17   1.09   1.25
## 20 Hospital_Stroke                     2.44   1.40   4.39
## 21 Hospital_TIA                        1.04   0.859  1.25
## 22 Hypertension                        0.582  0.555  0.609
<table>
<thead>
<tr>
<th>Variable</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>#23 Ischmeic Heart Disease</td>
<td>0.996  0.932  1.06</td>
</tr>
<tr>
<td>#24 Intellectual_disability</td>
<td>0.905  0.748  1.09</td>
</tr>
<tr>
<td>#25 Intercept</td>
<td>0.159  0.145  0.173</td>
</tr>
<tr>
<td>#26 Ischemic_stroke</td>
<td>0.607  0.340  1.04</td>
</tr>
<tr>
<td>#27 Mood_Anxiety</td>
<td>0.967  0.916  1.02</td>
</tr>
<tr>
<td>#28 Mood_disorders</td>
<td>0.994  0.958  1.03</td>
</tr>
<tr>
<td>#29 Multiple_Sclerosis</td>
<td>0.865  0.690  1.08</td>
</tr>
<tr>
<td>#30 Multiple_mental_illness</td>
<td>0.912  0.883  0.942</td>
</tr>
<tr>
<td>#31 Neurocognitivedisorders</td>
<td>1.11   1.06   1.17</td>
</tr>
<tr>
<td>#32 Neuroticrelated_disorders</td>
<td>0.864  0.837  0.893</td>
</tr>
<tr>
<td>#33 Opioid_use</td>
<td>4.02   3.86   4.20</td>
</tr>
<tr>
<td>#34 OAT</td>
<td>2.53   2.41   2.66</td>
</tr>
<tr>
<td>#35 Osteoarthritis</td>
<td>0.833  0.799  0.869</td>
</tr>
<tr>
<td>#36 Osteoporosis</td>
<td>0.711  0.657  0.768</td>
</tr>
<tr>
<td>#37 Osteomyelitis</td>
<td>0.939  0.866  1.02</td>
</tr>
<tr>
<td>#38 Otherpsychoactivedruguse</td>
<td>1.34   1.26   1.42</td>
</tr>
<tr>
<td>#39 Parkinsonism</td>
<td>0.919  0.733  1.14</td>
</tr>
<tr>
<td>#40 Personality_disorders</td>
<td>1.35   1.29   1.42</td>
</tr>
<tr>
<td>#41 Polysubstance</td>
<td>2.44   2.34   2.55</td>
</tr>
<tr>
<td>#42 Psychoticdisorders</td>
<td>1.13   1.08   1.19</td>
</tr>
<tr>
<td>#43 Rheumatoid_Arthritis</td>
<td>0.901  0.815  0.994</td>
</tr>
<tr>
<td>#44 Sedative_and_hypnoticuse</td>
<td>1.02   0.915  1.14</td>
</tr>
<tr>
<td>#45 Sepsis</td>
<td>1.62   1.54   1.70</td>
</tr>
<tr>
<td>#46 Stimulant_use</td>
<td>1.89   1.77   2.02</td>
</tr>
<tr>
<td>#47 Substancerelated_Disorders</td>
<td>10.6   10.2  11.0</td>
</tr>
<tr>
<td>#48 Tissue_infection</td>
<td>0.747  0.719  0.776</td>
</tr>
<tr>
<td>#49 Tobacco_use</td>
<td>0.493  0.464  0.524</td>
</tr>
</tbody>
</table>

```r
ggplot(oddsratio, aes (x = Odds_Ratio, y = reorder(as.factor(Variables), +Odds_Ratio), color)) + geom_point(stat = "identity", shape = 15)+ geom_hline(yintercept = 1, linetype = "dashed", color = "grey") + geom_errorbar(aes(xmin =SDE2.5, xmax = SDE975, width = 0.4 ))
```
#XGboost’s Shap Values

data1 <- fullset %>%
  select(-od_case)
dataX <- as.matrix(data1)

param_list <- list(objective = "reg:squarederror",
                     eta = 0.02,
                     max_depth = 10,
                     gamma = 0.01,
                     subsample = 0.95)

mod <- xgboost::xgboost(data = dataX,
                        label = as.matrix(fullset$od_case),
                        params = param_list,
                        nrounds = 10,
                        verbose = FALSE,
                        nthread = parallel::detectCores() - 2,
                        early_stopping_rounds = 8)

shap_values <- shap.values(xgb_model = mod, X_train = dataX)

shap_long <- shap.prep(shap_contrib = shap_values$shap_score, X_train = dataX)

shap.plot.summary(shap_long)
shap_int <- shap.prep.interaction(xgb_mod = mod, X_train = dataX)

shap_values$mean_shap_score

```r
##                            Opioiduse           substancerelated.disorders
##                         8.763785e-03                         7.172934e-03
##                        Polysubstance                                  ost
##                         2.938361e-03                         1.734264e-03
##                  hypertension_date_b                      Tissueinfection
##                         1.436867e-03                         1.285940e-03
##                           Stimulantuse                           Tobaccouse
##                         7.530721e-04                         4.585098e-04
##                           Cocaineuse                           Alcoholuse
##                         4.510072e-04                         4.198632e-04
##                               Sepsis                Multiplementalillness
##                         3.798863e-04                         3.650868e-04
##             Neuroticrelateddisorders                      mood_anx_date_b
```
### Variable importance from Boruta

```r
table(fullset$sod_case)
```

```r
## 0 1
## 446710 36679
```

```r
Borutamodel_v1 <- Boruta(od_case ~ ., data = fullset, doTrace = 2)
```

```r
saveRDS(Borutamodel, file = "borutamodel_V1.rds")
```

```r
plot(Borutamodel_v1)
```
**getSelectedAttributes** (Borutamodel_v1, withTentative = F)

```r
## Variable importance from caret
```

```r
## [1] "ost"                              "rheumatoid_arthritis_date_b"
## [3] "parkinsonism_date_b"               "osteo_arthritis_date_b"
## [5] "osteo_porosis_date_b"              "ms_date_b"
## [7] "mood_anx_date_b"                   "isch_stroke_date_b"
## [9] "ihd_date_b"                        "hypertension_date_b"
## [11] "hosp_tia_date_b"                   "hosp_stroke_date_b"
## [13] "heart_failure_date_b"             "haemorr_stroke_date_b"
## [15] "epilepsy_date_b"                  "diabetes_date_b"
## [17] "depression_date_b"                "copd_date_b"
## [19] "ckd_date_b"                       "asthma_date_b"
## [21] "angina_date_b"                    "ami_date_b"
## [23] "alzheimer_dementia_date_b"        "Tobaccouse"
## [25] "Tissueinfection"                  "substancerelated.disorders"
## [27] "Stimulantuse"                     "Sepsis"
## [29] "Sedativeandhypnoticuse"           "psychoticdisorders"
## [31] "Polysubstance"                    "Personalitydisorders"
## [33] "Otherpsychoactivedruguse"         "Osteomyelitis"
## [35] "Opioiduse"                        "Neuroticrelateddisorders"
## [37] "Neurocognitivedisorders"         "Multiplementalillness"
## [39] "Mooddisorders"                    "Intellectualdisability"
## [41] "Hallucinogensuse"                 "Endocarditis"
## [43] "Developmentdisorders"            "Cocaineuse"
## [45] "Cannabinoiduse"                   "Alcoholuse"
```
fullset <- fullset %>%
  mutate(od_case = recode(od_case, "0" = "No", "1" = "Yes"))

fullset$sod_case <- as.factor(fullset$sod_case)

rPartMod <- train(od_case ~ ., data = fullset, method = "rpart")

rpartImp <- varImp(rPartMod)

print(rpartImp)

## rpart variable importance
##
## only 20 most important variables shown (out of 48)
##
## Overall
## Polysubstance                        100.0000
## substancerelated.disorders            86.3302
## Opioiduse                             84.4518
## ost                                   62.1243
## Stimulantuse                          40.8137
## Otherpsychoactivedruguse              10.2704
## Cocaineuse                            9.1823
## hypertension_date_b                   3.6583
## alzheimer_dementia_date_b             0.6627
## hosp_stroke_date_b                    0.0000
## epilepsy_date_b                       0.0000
## angina_date_b                         0.0000
## heart_failure_date_b                  0.0000
## ckd_date_b                            0.0000
## Sedativeandhypnoticuse                0.0000
## Neuroticrelateddisorders              0.0000
## Behaviouralpsychologicaldisturbances   0.0000
## Tissueinfection                       0.0000
## copd_date_b                           0.0000
## Endocarditis                          0.0000

roc_imp <- filterVarImp(x = fullset[,1:48], y = fullset$sod_case)

roc_imp <- data.frame(cbind(variable = rownames(roc_imp), score = roc_imp[,1]))

roc_imp$score <- as.double(roc_imp$score)

library(dplyr)

rocscore <- roc_imp %>% arrange(desc(score))

print(rocscore, nrow(Inf))

##                                variable     score
## 1            substancerelated.disorders 0.8416637
## 2                             Opioiduse 0.7257841
## 3                                   ost 0.7123123
## 4                                   ost 0.6218935
library(randomForest)
rf = randomForest(x = fullset[,1:48], y = fullset$od_case)

var_imp_rf <- varImp(rf, scale = FALSE)
var_imp_rf1 <- var_imp_rf %>% arrange(desc(Overall))
view(var_imp_rf1, nrow(Inf))

##RFE Boosting Bagging to do feature selection

filterCtrl <- rfeControl(functions = treebagFuns, method = "cv", number = 10)

boostingrfe <- rfe(x = fullset[, 1:48], y = fullset$od_case, sizes = c(1:48), rfeControl = filterCtrl)

boostingrfe

## The top 5 variables (out of 10):
-- substancerelated.disorders, Polysubstance, Opioiduse, ost, Stimulantuse
```r
predictors(bootingrfe)
## [1] "substancerelated.disorders" "Polysubstance"
## [3] "Opioiduse"      "ost"
## [5] "Stimulantuse"   "hypertension_date_b"
## [7] "Tissueinfection" "Neuroticrelateddisorders"
## [9] "asthma_date_b"  "Mooddisorders"

boostingrfe$optVariables
## [1] "substancerelated.disorders" "Polysubstance"
## [3] "Opioiduse"      "ost"
## [5] "Stimulantuse"   "hypertension_date_b"
## [7] "Tissueinfection" "Neuroticrelateddisorders"
## [9] "asthma_date_b"  "Mooddisorders"

rfevarimp <- data.frame(varImp(bootingrfe))
rfevarimp

# Overall
## substancerelated.disorders 15423.0197
## Polysubstance              10431.9682
## Opioiduse                   9675.0254
## ost                         5468.9779
## Stimulantuse                3742.3126
## hypertension_date_b         1602.6666
## Tissueinfection             1416.2066
## Neuroticrelateddisorders    847.3272
```
```{r fig.width = 14, fig.height= 14}
library(corrplot)
M <- cor(fullset)
corrplot(M, method = "color")
```