2020-05-27

Predicting Death by Suicide with Administrative Health Care System Data

Sanderson, Michael

UNIVERSITY OF CALGARY

Predicting Death by Suicide with Administrative Health Care System Data

by

Michael Sanderson

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN COMMUNITY HEALTH SCIENCES

CALGARY, ALBERTA

MAY, 2020

ABSTRACT

Quantifying suicide risk with risk scales is common in clinical practice, but the performance of risk scales has been shown to be limited. Prediction models have been developed to quantify suicide risk and have been shown to outperform risk scales, but these models have not been commonly adopted in clinical practice. The original research presented in this thesis as three manuscripts evaluates the performance of prediction models that quantify suicide risk developed with administrative health care system data.

The first two manuscripts were designed to determine the most promising prediction model class and temporal data requirements. The modeling dataset contained 3548 persons that died by suicide and 35,480 persons that did not die by suicide between 2000 and 2016. 101 predictors were selected, and these were assembled for each of the 40 quarters prior to the quarter of death, resulting in 4040 predictors for each person. Logistic regression, feedforward neural network, recurrent neural network, one-dimensional convolutional neural network, and gradient boosted trees model classes were compared. The gradient boosted trees model class achieved the best performance and 8 quarters of data at most were required for optimal performance.

The third manuscript applied the findings from the first two manuscripts to evaluate the performance of prediction models in a clinical setting. The prediction models quantified the risk of death by suicide within 90 days following an Emergency Department visit for parasuicide. The modeling dataset contained 268 persons that died by suicide and 33,426 persons that did

not die by suicide between 2000 and 2017. The predictors were assembled for each of the 8 quarters prior to the quarter of death, resulting in 808 predictors for each person. Logistic regression and gradient boosted trees model classes were compared. The optimal gradient boosted trees model achieved promising discrimination and calibration.

Following the manuscripts, this thesis discusses further research. At present, there is no clinical consensus on the preferred performance characteristics for quantifying suicide risk. The critical next step for further research is to discover the preferred performance characteristics for quantifying suicide risk and to discover whether the preferred performance characteristics can be achieved.

PREFACE

The following three manuscripts comprise the original research supporting this thesis.

Sanderson M, Bulloch A, Wang J, Williamson T, Patten S. Predicting Death by Suicide Using Administrative Health Care System Data: Can Feedforward Neural Network Models Improve Upon Logistic Regression Models? Journal of Affective Disorders. 2019; 257:741-747.

Sanderson M, Bulloch A, Wang J, Williamson T, Patten S. Predicting Death by Suicide Using Administrative Health Care System Data: Can Recurrent Neural Network, One-Dimensional Convolutional Neural Network, and Gradient Boosted Trees Models Improve Prediction Performance? Journal of Affective Disorders. 2020; 264:107-114.

Sanderson M, Bulloch A, Wang J, Williams KG, Williamson T, Patten S. Predicting Death by Suicide Following an Emergency Department Visit for Parasuicide With Administrative Health Care System Data and Gradient Boosted Trees. EClinicalMedicine. 20 (2020) 100281.

The first author curated the data, conducted the analyses, interpreted the results, and wrote the manuscripts. All authors critically revised and contributed intellectually to the manuscripts. The manuscripts are reproduced in their entirety as Chapters 2, 3, and 4 in this thesis. Reproduction of the manuscripts in this thesis complies with Elsevier policies. Permission to include the above manuscripts in this thesis was obtained from all co-authors.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Background

Suicide is a leading cause of death in Canada [1] and internationally [2]. Over half of all deaths by suicide in Alberta between 2000 and 2017 occurred in persons under 45, and 96 percent occurred in persons under 75, resulting in 290,490 years of life lost [3].

The suicide rate in Alberta declined between 1983 and 2014, but increased sharply in 2015 (see: Figure 1). The decrease in the overall suicide rate has largely been attributable to a decrease in the male suicide rate (see: Figure 2). The decrease coincides with the introduction of Fluoxetine (Prozac) in the late 1980s for the treatment of major depression and the decrease may also coincide with a decrease in mental health stigma and a corresponding increase in seeking treatment. The male suicide rate was 3.5 times higher than the female suicide rate between 1983 and 2017, although this difference has been decreasing over time (see: Figure 2). The difference between the male suicide rate and the female suicide rate was lowest in the teenage years and the mid-thirties to late-fifties, while the difference was highest in the twenties to mid-thirties and over 60 (see: Figure 3). The male suicide rate in persons over 90 was especially high (42 per 100,000) and accounted for 0.6% of all male deaths by suicide. Mental health likely plays a larger role in younger persons and physical health likely plays a larger role in older persons.

The suicide rate between 2000 and 2017 in Alberta tended to be higher in rural communities than in urban communities (see: Figures 4 and 5). Communities with higher Low-Income Measure After Tax (LIM-AT, a measure of after-tax household income adjusted for household size) and higher unemployment rates tended to have higher suicide rates (see: Figures 6 and 7). Rurality itself may not be an independent risk factor for suicide and it may be that rural communities have higher suicide rates because they also tend to have lower access to mental health supports, higher LIM-AT, and unemployment rates.

Of the persons that died by suicide in Alberta in 2017, 87 percent had visited a physician in the year prior to the date of suicide and 68 percent had received a mental health diagnosis [4]. In the 90 days prior to the date of suicide, 70 percent had visited a physician and 59 percent had received a mental health diagnosis [4].

1.2 Objective

Death by suicide is an event where health care service and health care policy interventions are focused entirely on prevention, since death by suicide cannot be treated. Health care service providers and health care policy providers must be able to quantify the risk of death by suicide in order to choose the optimal prevention intervention. This is because there are a number of clinical and population interventions that can be chosen to prevent death by suicide in a particular setting, and they may vary in intensity, invasiveness, effectiveness, expense, and suitability. In health care service settings, clinical judgment and risk scales are most commonly

relied upon to quantify suicide risk but these have been shown to have low to moderate prediction performance (see: section 1.3. below).

There is an opportunity to evaluate the performance of prediction models for quantifying suicide risk developed with administrative health care system data and machine learning model classes. Administrative health care system data may be valuable for developing prediction models because of the volume and breadth of data, and because its ongoing collection means that prediction models developed with this data can be used in the future. Machine learning model classes may be valuable for developing prediction models because of their ability to learn complex non-linear relationships, and because machine learning software and hardware has become available to non-specialists.

The objective of this thesis is to evaluate the performance of prediction model classes that quantify suicide risk with predictors available in electronic administrative health care system data. In principle, the prediction model classes, administrative health care system data, and evaluation methods reported in this thesis could be used to inform both clinical and population interventions. However, the focus of this thesis will be on the development and evaluation of prediction models for clinical settings. For example, as described above, 70 percent of persons that died by suicide in 2017 had visited a physician in the 90 days prior to the date of suicide. For many of these persons, the physician may not have been aware of the high risk of suicide. Even if the physician was aware, the physician may not have known the precise risk, and knowing the precise risk may have led to a better choice of intervention.

This thesis contains three manuscripts. The first two manuscripts used a case-control study design in order to include all available instances of death by suicide in the modeling dataset, and were designed to discover the most promising prediction model class from among logistic regression, feedforward neural networks, one-dimensional convolutional neural networks, recurrent neural networks, and gradient boosted trees. These manuscripts focused on the ability of prediction models to discriminate between persons at lower and higher risk of death by suicide. The third manuscript used a retrospective open-cohort study design, and was designed to evaluate the performance of prediction models in a realistic health care setting (emergency department visits for parasuicide) with the most promising model class from the first two manuscripts.

1.3 Literature Review

A literature review was carried out for this thesis. The goals of the literature review were to identify predictors of suicidality risk, to identify and evaluate existing approaches that quantify suicidality risk, and to discover whether there is an opportunity for prediction models that quantify suicide risk developed with administrative health care system data and machine learning model classes to provide novel and needed contributions. Suicidality is comprised of suicide, parasuicide, and suicidal ideation. It is important to note that suicide, parasuicide, and suicidal ideation are related but distinct behaviors. Suicide is the intentional death of oneself. The term ‘parasuicide’ can have different meanings in the literature, but in this literature review, parasuicide is suicidal behaviour that did not result in death. The reason the term ‘parasuicide’ is used in this literature review is that determining intent to die is difficult and

behaviour commonly labeled ‘attempted suicide’ is not necessarily an attempt at suicide. In this literature review, suicidal ideation is personal thoughts of suicide or suicidal behaviours.

Any study title that was related to suicidality (suicide, parasuicide, suicidal ideation), and where the predictors were similar to those available for this thesis or where the focus was on prediction methods for suicidality, was selected. Studies with a military population or studies that were focused on genetic predictors were excluded because they were not relevant to this thesis. The abstracts of 53 studies were reviewed (by the author), and 23 studies that were directly related to the goals of the literature review were selected for full review. Those 23 studies were categorized according to outcome (some studies examined more than one outcome): suicide (10), parasuicide (15), or suicidal ideation (3); at-risk population: general population (4) or sub-population (19); risk factor measurement: administrative/survey data (8) or clinical assessment tool (11) or a summary of other work (4). Appendix C contains an evidence table summarizing each study selected for full review.

1.3.1 Literature Review: Risk Factors for Suicidality

This section describes studies that included suicide, parasuicide, and suicidal ideation. The majority of risk factors for suicidality in the studies selected for full review were measured using clinical assessment tools (11), either developed for estimating the risk of suicidality in clinical practice or developed specifically for that study. Although the clinical assessment tools differed across the studies, the risk factors for predicting suicidality tended to be consistent. The most common and strongest risk factors for predicting suicidality using clinical assessment tools were

parasuicide, suicidal ideation, and the presence of a mental disorder. While a mental disorder can be classified in different ways in different studies, the Diagnostic and Statistical Manual of Mental Disorders (DSM–5) describes a mental disorder as “a syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behavior that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning.”. Other risk factors included poor social conditions and poor social interactions, age and sex, lethality of parasuicide, intensity of suicidal ideation, substance misuse, health care provider suicidality risk assessment, and physical illness.

The remaining studies (8) collected risk factors for suicidality from administrative or survey data that were not specially designed to predict suicidality risk. The most common and strongest risk factors for predicting suicidality using administrative data were parasuicide and the presence of mental illness. Other risk factors described were generally similar to those measured by clinical assessment tools as above but also included different perspectives. For example, one study that used administrative data [5] described ecological risk factors for suicide and parasuicide in subway stations. The study largely examined the structural characteristics of the train stations themselves and the structural characteristics of the surrounding areas, and not population characteristics in detail.

The summary studies (4) reiterated the risk factors for suicidality from the literature, and then discussed strategies for improving and applying the estimates of suicidality risk towards suicide prevention. For example, Pisani et al. [6], recommended that psychiatrists-in-training be taught

more prevention-oriented suicidality risk formulations that include: an individual's stratified risk relative to others, an individual's risk state relative to previous personal risk states, available crisis resources, and foreseeable changes that may exacerbate risk. The summary studies all noted that interventions require being able quantifying suicide risk, but that but that quantifying suicide risk is difficult in practice.

1.3.2 Literature Review: Prediction Modeling

The studies by Wang et al. [7], Tran et al. [8], and Yaseen et al. [9] described below reported the receiver operating characteristic curve (AUC) while most of the other studies selected for full review reported effect sizes and estimates of statistical significance from regression models. While effect sizes and estimates of statistical significance are important for estimation modeling, they do not evaluate prediction performance. This is an important distinction because the goal of estimation modeling is to understand how risk factors contribute to the risk of suicide, while the goal of prediction modeling is strictly prediction performance. Estimation modelers may prefer to exchange prediction performance for a better understanding of the relationship between risk factors and suicide.

Wang et al. [7] used logistic regression to predict emergency department (ED) visits for parasuicide within 6 months in persons that had been referred for psychiatric services in the EDs of two large tertiary care hospitals in Manitoba (n = 2792). There were 2792 persons and 136 (5 percent) visited an ED for parasuicide within 6 months. The authors found that senior psychiatric residents (AUC = 0.76) and staff psychiatrists (AUC = 0.78) were better able to

predict than non-psychiatric residents (AUC = 0.59), junior psychiatric residents (AUC = 0.68), and a clinical assessment tool (AUC = 0.62). This suggests that psychiatric residents and staff psychiatrists were better than non-psychiatric residents, junior psychiatric residents, and a clinical assessment tool at distinguishing between persons that visited an ED for parasuicide within 6 months and persons that did not.

Tran et al. [8] used ordinal regression to predict parasuicide within 180 days in persons that had undergone a suicide risk assessment (n = 7399) at the only tertiary hospital or one of five community health centres in the Barwon Health regional health service in Australia. The authors compared clinician risk assessments using a risk assessment checklist with an ordinal regression model using administrative hospital data to predict parasuicide at 30, 60, 90, and 180 days after the suicide risk assessment. Parasuicide was classified into three groups: low risk (no parasuicide, or an ED or inpatient admission with low-lethality parasuicide), moderate risk (an ED or inpatient admission with moderate-lethality parasuicide), and high risk (an ED or inpatient admission with high-lethality parasuicide). At 30, 60, 90, and 180 days after the suicide risk assessment, clinicians predicted parasuicide in high risk versus moderate and low risk combined with AUC between 0.55 and 0.59, while the prediction model predicted with AUC between 0.73 and 0.79. At 30, 60, 90, and 180 days after the suicide risk assessment, clinicians predicted parasuicide in high and moderate risk combined versus low risk with AUC between 0.52 and 0.54, while the prediction model predicted with AUC between 0.71 and 0.79.

Yaseen et al. [9] used a clinical assessment tool to predict parasuicide in persons that were psychiatric inpatients in two tertiary care hospitals in New York City after being seen in the ED with parasuicide or suicidal ideation (n = 161). The authors used a clinical assessment tool called the 'Suicide Trigger Scale v.3' (STS-3) designed to measure a "suicide trigger state" thought to precede parasuicide. Although the authors recruited 161 participants, they were only able to contact 54 for follow-up. Of the 54 persons that were contacted, 13 (24 percent) reported parasuicide. The original STS-3 did not predict parasuicide, but a transformed STS-3 score (AUC = 0.73) and a subset of the STS-3 scores (AUC = 0.81) that were created based on a post hoc analysis did. Although the discrimination performance of the post-hoc STS-3 scores could be considered promising, the performance was not validated and so is likely overly optimistic.

Karmakar et al. [10] used risk stratification (not modeling) and electronic hospital records to predict the risk of parasuicide in mental health patients in the same health region as Tran et al. (2014) in order to create a risk score that could be used as a decision support tool by hospital clinicians. The risk score was a stratification of parasuicide risk based on physical health diagnoses codes, rather than mental health diagnosis codes. The risk score outperformed routine risk assessment (AUC = 0.56) but only demonstrated moderate performance (AUC = 0.71).

There were three studies that used machine learning model classes to quantify suicidality risk. None of the three studies used performance validation methods. Bae et al. [11] used decision tree

analysis with a national mental health survey of middle and high school students to predict parasuicide. Prediction performance was not reported but the authors used the proportion of parasuicide in each tree node to measure the relative importance of predictors (a high proportion of parasuicide in a tree node indicates high importance). The most important predictor was severity of depression. Other important predictors included delinquency, intimacy with family, and stress. Poulin et al. [12], used genetic programming to predict the risk of suicide using the text of clinical notes in U.S. Veterans Administration medical records. Keywords and multi-word phrases were used to distinguish between equal-sized groups of veterans that had and had not died by suicide, such as "agitation" (persons that died by suicide), "disheveled" (persons that did not die by suicide but received psychiatric treatment), and "plasma" (persons that did not die by suicide and did not receive psychiatric treatment). The authors evaluated the prediction accuracy of the presence of single words, word pairs, word triples, or phrases. The authors reported mean accuracies between 46-65 percent. The goal of the study was not to develop prediction models but to investigate whether single words and multi-word phrases could demonstrate prediction utility. Cook et al. [13], used Natural Language Processing (NLP) with respondents' unstructured text responses to the question "How are you feeling today?" (in Spanish), to predict suicidal ideation in adults discharged from psychiatric treatment in a hospital. The authors compared the ability of a NLP model with the response to the question "How are you feeling today?" as the predictor, and logistic regression with tabular data from a survey as predictors, to predict reported suicidal ideation. The authors found that the NLP model achieved a sensitivity of 0.56, specificity of 0.57, and PPV of 0.61, while the logistic regression model achieved a sensitivity of 0.76, specificity of 0.62, and PPV of

0.73. Although the PPVs appear promising, the prevalence of suicidal ideation in the modeling dataset was 53 percent.

1.3.3 Literature Review: Summary

The studies above showed that there is generally consensus regarding the risk factors that are important for quantifying suicidality risk. However, most studies focused on estimation modeling rather than prediction modeling, and the studies that evaluated prediction performance reported low to moderate performance. The volume of data in the studies above tended to be small, and assembled from a single data source. The studies also showed that quantifying suicidality risk is difficult, and several noted that suicide was especially difficult due to its rarity.

The studies that used large administrative datasets achieved the best prediction performance among the studies in the literature review, but they did not use machine learning model classes. The studies that used machine learning model classes did not use large administrative datasets. It seems from the literature review that there is an opportunity to evaluate the performance of prediction models that quantify suicide risk developed with administrative health care system data and machine learning model classes. The administrative health care system data available for this thesis was larger and more comprehensive than those reported in the literature review, and machine learning software, hardware, and model classes have advanced significantly compared to what was available when the studies in the literature review were completed. Thus, while the results of this thesis cannot be directly compared to

earlier prediction models for the reasons described above, this thesis can provide novel and needed contributions to discover whether prediction performance can be achieved that could lead to clinical applications. This thesis did not seek to develop a finalized prediction model for comparison with prior studies but instead sought to take the first steps towards developing prediction models for clinical practice and to describe best practices for developing prediction models for clinical practice.

CHAPTER 2: PREDICTING DEATH BY SUICIDE USING ADMINISTRATIVE HEALTH CARE SYSTEM DATA: CAN FEEDFORWARD NEURAL NETWORK MODELS IMPROVE UPON LOGISTIC REGRESSION MODELS?

This manuscript was published in the Journal of Affective Disorders in October, 2019. This manuscript was designed to discover the prediction utility inherent in the administrative health care system data in Alberta compared with the studies in the literature review, and to evaluate the relative performance of logistic regression and a classic machine learning model class.

2.1 Abstract

Background

Suicide is a leading cause of death worldwide. With the increasing volume of administrative health care data, there is an opportunity to evaluate whether machine learning models can improve upon statistical models for quantifying suicide risk.

Objective

To compare the relative performance of logistic regression and single hidden layer feedforward neural network models that quantify suicide risk with predictors available in administrative health care system data.

Methods

The modeling dataset contained 3548 persons that died by suicide and 35,480 persons that did not die by suicide between 2000 and 2016. 101 predictors were selected, and these were assembled for each of the 40 quarters (10 years) prior to the quarter of death, resulting in 4040 predictors in total for each person. Logistic regression and single hidden layer feedforward neural network model configurations were evaluated using 10-fold cross-validation.

Results

The optimal feedforward neural network model configuration (AUC: 0.8352) outperformed logistic regression (AUC: 0.8179).

Limitations

Many important predictors are not available in administrative data and this likely places a limit on how well prediction models developed with administrative data can perform.

Conclusions

Although the models developed in this study showed promise, further research is needed to determine the performance limits of statistical and machine learning models that quantify suicide risk, and to develop prediction models optimized for implementation in clinical settings.

2.2 Introduction

Suicide is a leading cause of death worldwide [1]. Between 2000 and 2017 in Alberta, Canada, suicide accounted for 23 percent of all deaths among persons 15 to 30 and 16 percent of all deaths among persons 30 to 45 [2]. Over the same time period, 96 percent of deaths by suicide in Alberta occurred in persons under 75 causing 289,078 years of life lost [2]. Suicide also accounted for two percent of all deaths in persons over nine and ten percent of person years of life lost [2].

For health care service providers and health care policy providers to take actions to reduce suicide risk, they must be able to quantify suicide risk. Unfortunately, quantifying suicide risk is difficult [3, 4] because suicide is rare and the risk factors for suicide are common. Further, many risk factors for suicide do not vary over time which makes acute quantification of suicide risk even more difficult. Attempts to quantify suicide risk with statistical models have predicted suicide better than chance, but have not achieved performance sufficient to be broadly useful [5, 6, 7, 8]. Machine learning models such as neural networks have achieved successes with many difficult prediction problems and this has led to discussion about whether similar success could be achieved with suicide prediction [9].

Administrative health care data is potentially valuable for developing prediction models because of the volume and breadth of data, and because its ongoing collection means that prediction models developed with this data can be used in the future to inform health care services and policies. With the availability of machine learning hardware and software to non-specialists, and with the increasing volume of administrative health care data, there is an

opportunity to evaluate whether feedforward neural network models can improve upon logistic regression models for quantifying suicide risk.

2.3 Objective

The objective of this study is to compare the relative performance of logistic regression and single hidden layer feedforward neural network (FNN) models that quantify suicide risk with predictors available in administrative health care system data as a first step towards evaluating whether machine learning is a promising avenue of research in this domain. The objective is not to develop optimized models for implementation in clinical settings or to identify important predictors but rather to investigate whether FNN models are capable of providing an improvement in prediction performance compared with logistic regression models using identical modeling datasets.

If FNN models prove capable of outperforming logistic regression models, then FNN models would be promising for future research to develop optimized models for implementation in health care service and policy settings. For example, a computer system could be developed that would assemble all of the administrative records for a person that is about to be discharged from a psychiatric inpatient setting and then use an optimized FNN model to estimate of the risk of death by suicide in the near term. A clinician could use that risk estimate when developing a discharge care plan.

2.4 Neural Networks

Neural networks are a flexible class of machine learning models that were inspired by neuroscience [2]. Conceptually, a neural network model is made up of layers of neurons, with the neurons in one layer connected to the neurons the next. Each neuron is a computational unit that multiplies its input values by a corresponding set of learnable weight parameters, sums the multiplied values, transforms the summed value using a nonlinear activation function, and outputs the transformed value.

The first layer in a neural network model is the input layer, and each unit in the input layer contains the value of one of the predictors for a particular observation. The input layer passes all predictor values for a particular observation to each neuron in the first hidden layer. Each neuron in the first hidden layer computes a different function with the predictor values. The first hidden layer then passes its output values to each neuron in the second hidden layer, where each neuron computes a different function with its input values, and so on to the final output layer which makes a prediction.

Neural network models learn by iteratively comparing its predictions with the observed outcomes and then updating its weight parameters to improve its predictions. Neural network models have a number of hyperparameters that are set by the modeler, including the number of neurons in each hidden layer, the number of hidden layers, the learning rate, and the number of epochs. The learning rate is how much the weight parameters are adjusted at each

iteration, and the number of epochs is the number of times the entire training dataset is used to update the weight parameters.

2.5 Methods

2.5.1 Data Sources

The population in this study was the province of Alberta in Canada. Alberta has a publicly-funded single-payer health care system with administrative data systems that record the health care services of nearly its entire population of 4.07 million people [10]. The outcome and predictors were selected from administrative health care data in Alberta.

The outcome, death by suicide, was obtained from the Alberta Vital Statistics Cause of Death data system (ICD-10 cause of death codes X60 through X84). Predictors were assembled from the following data systems: the Alberta Health Care Insurance Plan (AHCIP) Registry, Supplemental Enhanced Service Event (SESE; physician claims), Morbidity and Ambulatory Care Abstract Reporting (MACAR; ambulatory care and inpatient hospitalizations), Pharmaceutical Information Network (PIN; community pharmacy dispenses), and the Alberta Disease Registry for Surveillance (a registry containing the date Albertans met disease case definitions). The AHCIP Registry records the residency status of Albertans each quarter, and so the other datasets were assembled by quarter to match the temporal granularity of the residency data. The datasets were linked for this study using the unique Personal Health Number assigned to all Albertans for the delivery of health care services. This study was approved by the University of Calgary Conjoint Health Research Ethics Review Board.

2.5.2 Hardware and Software

The administrative data were extracted and assembled using SAS 9.4. The analysis was performed on a desktop computer with an Ubuntu 18.04.1 LTS operating system and a GeForce GTX 1080 Ti 12GB graphics processing unit (GPU) using the NVIDIA-SMI 390.87 driver. The analysis was written in the Python programming language in a Jupyter 5.6.0 notebook in Anaconda Navigator 1.8.7. The logistic regression models were developed using scikit-learn 0.20.0. The FNN models were developed with Keras 2.2.2 using the TensorFlow backend with GPU support.

2.5.3 Inclusion and Exclusion Criteria

It was not computationally feasible to include all persons in the administrative health care data in the modeling dataset. For each person that died by suicide in Alberta between 2000 and 2016 (3548), 10 persons that did not die by suicide and were residing in Alberta in the quarter of death were randomly selected (35,480) using the proc surveyselect function in SAS 9.4. This ratio was chosen to generate a modeling dataset large enough to produce robust models while also being computationally feasible on a desktop computer with a GPU. Residents of Alberta 10 years and older were included, as 10 years is the age when suicide risk begins to manifest in the administrative data [2]. No other inclusion, exclusion, or matching criteria were applied.

2.5.5 Outcome Class Weights

As described above, 10 persons that did not die by suicide were randomly selected for each person that died by suicide, and so the outcome class distribution was imbalanced. In order to

assign equal importance to both outcome classes, the models included class weights of 10 / 11 for persons that died by suicide and 1 / 11 for persons that did not die by suicide.

2.5.6 Predictors

Predictors were selected from the administrative health care data based on those identified as having suicide or parasuicide prediction utility in a literature review carried out for this study. 101 predictors were selected, and these were assembled for each of the 40 quarters (10 years) prior to the quarter of death, resulting in 4040 predictors in total (101 predictors x 40 quarters) for each person. Generally, the predictors selected were related to mental health, but predictors related to physical health were also selected because they have been shown to have utility for suicide prediction [11]. Although some of the predictors related to physical health may not appear to be directly related to suicide, they were included in the modeling dataset to allow the models the opportunity to learn complex relationships. For example, gout is likely not directly related to suicide but combined with other predictors related to physical health, gout could contribute to the overall burden of physical illness.

A full listing of the selected predictors is available in Appendix B. The modeling dataset contained 39,028 rows (3548 persons that died by suicide and 35,480 persons that did not die by suicide) and 4041 columns (4040 predictors and 1 outcome).

2.5.7 Missing Values

Missing predictor values were only present due to a person not being resident in Alberta during a particular quarter, and these were assigned a value of zero. To distinguish missing predictor values from true zeroes, a residency flag was included as a predictor to indicate whether each person was resident in Alberta during a particular quarter.

2.5.8 Model Configuration Evaluation

A model is a single realization of a model configuration. For example, a model configuration might be written as $Y = B_0 + B_1(X_1)$ while a model developed with a particular modeling dataset might be written as $Y = 15 + 0.01(X_1)$. Machine learning model configurations do not lend themselves to hypothesis tests and confidence intervals, and are instead commonly evaluated empirically with a validation dataset [12]. A modeling dataset is randomly divided into a training dataset and a validation dataset, and then a model is developed with the training dataset and evaluated with the validation dataset [1]2. The validation dataset is used to provide an estimate of the expected performance of the model configuration with data the model was not developed with (unseen data).

A problem with using a single validation dataset to evaluate a model configuration is that there are many possible random divisions of the modeling dataset into training and validation datasets. Training and validation datasets will vary from one random division of the modeling dataset to the next, and so the resulting models and validation estimates will also vary. The same model configuration developed with one random division of the modeling dataset into

training and validation datasets will result in a different – but hopefully very similar – model than another random division.

K-fold cross-validation is an evaluation approach that uses k validation datasets to obtain a more robust estimate of expected performance with unseen data than with a single validation dataset [12]. First, the modeling dataset is randomly divided into k approximately equally-sized parts (k = 5 or 10 is common). Then, a model is developed with k – 1 parts acting as a training dataset and evaluated with the remaining part acting as a validation dataset; this process is repeated until all k parts have acted as a validation dataset once [12]. The mean of the k validation estimates is the k-fold cross-validation estimate of the expected performance of the model configuration with unseen data [12].

The k-fold cross-validation receiver operating characteristic area under the curve (AUC) was chosen as the single metric to evaluate model configuration performance because it has the intuitive interpretation that the AUC is the probability that the predicted risk was higher for a person that died by suicide than a person that did not [13], and because it was closely associated with sensitivity, specificity, positive prediction value (PPV), and negative prediction value (NPV).

2.5.9 FNN Model Configurations

FNN model configurations have hyperparameters to be tuned. The hyperparameters evaluated in this study were the number of neurons in the hidden layer (8, 16, 32, 64, 128), the learning rate (1e-4, 5e-5, 1e-5, 5e-6, 1e-6), and the number of epochs (50, 100, 150, 200, 250, 300, 350,

400, 450, 500). Each combination of the above hyperparameter settings were evaluated, resulting in 250 FNN model configurations. Batch size is a hyperparameter, but this study defaulted to a batch size of 512. The hidden layer activation function is also a hyperparameter, but this study defaulted to the Rectified Linear Unit (ReLU) activation function.

Each of the 250 FNN model configurations and the single logistic regression model configuration were evaluated with 10-fold cross-validation. The evaluation took approximately 3 days of compute time (251 models x 10 folds per model = 2510 models in total).

2.6 Results

2.6.1 Performance Metrics

The 10-fold cross-validation AUC estimate for logistic regression was 0.8179. The 10-fold cross-validation AUC estimate for the optimal FNN model configuration was 0.8352. The optimal FNN model configuration had 32 neurons, learning rate of 1e-5, and 300 epochs. Each FNN neuron configuration was capable of achieving essentially the same maximum 10-fold cross-validation AUC, although the learning rate and epoch combination required to achieve the maximum 10-fold cross-validation AUC varied. The neuron configuration with 32 neurons had the highest 10-fold cross-validation AUC but the other neuron configurations achieved a nearly identical optimal 10-fold cross-validation AUC (8: 0.8332, 16: 0.8336, 64: 0.8344, 128: 0.8338) and are all likely equally good estimates of the expected optimal AUC with unseen data.

In addition to the 10-fold cross-validation AUC, a number of other 10-fold cross-validation performance metrics were computed and are included in Table 1. The optimal FNN model had a greater sensitivity (0.6996) than the logistic regression model (0.6531), with similar specificity (0.8098, 0.8265), PPV (0.2961, 0.2734), and NPV (0.9642, 0.9597).

2.6.2 FNN Performance Trajectories

Each FNN neuron configuration (8, 16, 32, 64, 128) had the same performance trajectory across learning rates up to the maximum 10-fold cross-validation AUC of around 0.8350. Once the maximum AUC was reached, FNN configurations with more neurons overfit to the training data more severely. For example, in Figure 1, the maximum 10-fold cross-validation AUC was achieved for 8 neurons with a learning rate of 5E-5 and 150 epochs, although there were other configurations that achieved a nearly identical 10-fold cross-validation AUC.

With a learning rate of 1E-6 or 5E-6 or 1E-5, the maximum number of epochs in the evaluation (500) was not sufficient to achieve the maximum 10-fold cross-validation AUC, although the trajectories suggest that these configurations would eventually achieve the maximum 10-fold cross-validation AUC with enough epochs.

2.7 Discussion

The objective of this study is to compare the relative performance of logistic regression and single hidden layer FNN models that quantify suicide risk with predictors available in administrative health care system data. We sought to evaluate relative performance in the

simplest case, without model tuning techniques such as variable reduction or regularization. The optimal FNN model configuration (AUC: 0.8352) outperformed logistic regression (AUC: 0.8179), showing that FNN models can improve upon logistic regression models. FNNs appear to be promising for future research to develop optimized models for implementation in health care service and policy settings.

The AUCs of the logistic regression and optimal FNN models in this study were higher than those in other studies that predicted suicidality, possibly because the controls were sampled from the general population. In a study of parasuicide in persons that had been referred for psychiatric services in the Emergency Departments of two large tertiary care hospitals in Manitoba, Canada [14], senior psychiatric residents (AUC: 0.76) and staff psychiatrists (AUC: 0.78) were better able to predict parasuicide than non-psychiatric residents (AUC: 0.59), junior psychiatric residents (AUC: 0.68), and a clinical assessment tool (AUC: 0.62). In a study of parasuicide in Melbourne, Australia [15], prediction models developed with administrative hospital data (AUC: 0.71 to 0.79) were better able to predict parasuicide than clinicians (AUC: 0.52 to 0.54). In another study of parasuicide in Melbourne, Australia [16], a prediction model developed with administrative hospital data based on physical health diagnoses codes (AUC: 0.71) was better able to predict parasuicide than routine risk assessment (AUC: 0.56).

Further research is needed to determine the performance limits of statistical and machine learning models that quantify suicide risk. For example, predictor reduction techniques were not employed in this study because these approaches often require judgment, which would

preclude direct performance comparisons, but it is possible that fewer predictors or even composite predictors could result in improved prediction performance. It seems very possible that improved prediction performance could be achieved with more predictor engineering, more complex models, more data volume, and more suicide-specific predictors, but it is unclear how large the improvement in prediction performance might be.

Further research is needed to develop prediction models optimized for implementation in particular settings, such as arrival at an emergency department, discharge from a psychiatric inpatient stay, or policy development. The development, evaluation, and implementation of suicide risk quantification models for health care service providers and health care policy providers will require a more in-depth consideration of the implications of the performance metrics, particularly the impact of different risk distributions in different settings. Further, considerations beyond the performance characteristics of the models must also be considered. For example, one would have to consider the possibility of unintended consequences of assigning risk, such as the possibility that assigning risk, even if accurate and reliable, may do more harm than good if the assigned risk lead to inappropriate care.

Most importantly, further research is needed to determine whether prediction models can be developed that will be useful to health care service providers and health care policy providers. Prediction models outperform clinicians when predicting the risk of suicidality [11, 15, 16], but these models have not been widely implemented in clinical settings. Risk scales are commonly used in clinical settings, but they have limited utility for quantifying suicidality risk [5, 17, 18, 19, 20]. If

prediction models developed with administrative data in Alberta can be used in clinical settings, then similar prediction models could be developed and implemented across Canada because most Canadian provinces collect administrative data similar to that used in this study.

2.8 Limitations

2.8.1 Case-Control Sampling Design

Although the case-control sampling design used in this study is useful for evaluating model discrimination, it is not useful for evaluating model calibration because the risk distribution in the modeling dataset is not representative of any population or setting. That said, the objective of this study was not to develop a prediction model for implementation in a particular setting, but rather to explore the relative performance of logistic regression and FNN models that quantify suicide risk.

2.8.2 Administrative Data

Although being able to quantify suicide risk using data that will continue to be collected into the foreseeable future has benefits, the predictors available in the administrative data were not collected for the purposes of suicide risk quantification. Thus, many important predictors are not available in administrative data and this likely places a limit on how well prediction models developed with administrative data can perform. For example, while the administrative data used in this study contained proxies of severity of mental illness in terms of service utilization, severity itself was not directly measured. However, as electronic health data becomes richer

and the opportunity to link with non-health data grows, prediction models developed with linked electronic data may be even more promising.

2.8.3 Big (Enough?) Data

The modeling dataset in this study included all persons that died by suicide in Alberta between 2000 and 2016 (3548), but only included 35,480 persons that did not die by suicide due to the computational limitations of a desktop computer with a GPU. In general, machine learning models outperform statistical models when developed with large volumes of data. While the modeling dataset in this study might be considered 'big data' by suicide literature standards, it may not be a large enough volume for FNN models to learn a substantially more complex function than the logistic regression function. For example, the Large Scale Visual Recognition Challenge 2017 [21] used a training dataset with 1.2 million images.

2.8.4 Neural Networks and Deep Learning

FNN models were chosen for comparison with logistic regression models in this study because they are capable of learning very complex and non-linear relationships between predictors and outcomes, and because they scale well to large datasets. A drawback of FNN models is that they are essentially uninterpretable and so any improvement in prediction performance compared to logistic regression comes with the cost of opacity.

Deep learning neural network models are capable of reducing the overall number of parameters required to represent a function by adding hidden layers [22]. Although deep learning

models are often required to achieve optimal performance in many machine learning projects, only FNN models with a single hidden layer were explored in this study because the FNN models with higher capacity did not outperform the FNN models with lower capacity. As discussed above, each neuron configuration appeared to be able to achieve the maximum 10-fold cross-validation AUC and models with higher capacity began to overfit once the maximum was achieved.

2.8.5 Temporality

The prevention framework suggested by Pisani et al. [23] considers suicide risk to have two components: risk status (risk relative to other persons) and risk state (risk relative to prior personal states). Temporal precision is particularly important with suicide because risk can escalate to crisis in a very short period of time, and being able to predict escalating risk is crucial for health care service providers. The models developed in this study attempted to include both components when quantifying suicide risk. Although time was not included as a tensor dimension, a FNN with sufficient capacity (neurons) is capable of representing the sequential importance (risk state) of predictors, and indeed, any function [22, 24].

2.9 References

1. Preventing Suicide: A Global Imperative. World Health Organization. 2014. http://apps.who.int/iris/bitstream/10665/131056/1/9789241564779_eng.pdf?ua=1&ua=1

2. Alberta Vital Statistics. Cause of Death database; ICD-10: X60 through X84. 2019.

3. Mulder R, Newton-Howes G, Coid JW. The futility of risk prediction in psychiatry. Br J Psychiatry 2016; 209:271-272.

4. Large M, Kaneson M, Myles N, Myles H, Gunaratne P, Ryan C. Meta-analysis of longitudinal cohort studies of suicide risk assessment among psychiatric patients: heterogeneity in results and lack of improvement over time. PLoS ONE. 2016; 11(6):e0156322.

5. Chan MK, Bhatti H, Meader N, Stockton S, Evans J, O'Connor RC, et al. Predicting suicide following self-harm: systematic review of risk factors and risk scales. Br J Psychiatry. 2016; 209(4):277-83.

6. Huang X, Ribiero JD, Musacchio KM, Franklin JC. Demographics as predictors of suicidal thoughts and behaviors: A meta-analysis. PLoS ONE. 2017; 12(7):e0180793.

7. Carter G, Milner A, McGill K, Pirkis J, Kapur N, Spittal MJ. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. Br J Psychiatry. 2017.

8. Ribeiro JD, Franklin JC, Fox KR, Bentley KH, Kleiman EM, Chang BP, Nock MK. Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. Psychological Medicine. 2016; 46:225–236.

9. Ribeiro JD, Franklin JC, Fox KR, Bentley KH, Kleiman EM, Chang BP, Nock MK. Suicide as a complex classification problem: machine learning and related techniques can advance suicide

prediction – a reply to Roaldset. Psychological Medicine. 2016; 46:2009–2010.

10. Alberta Health: Overview of Administrative Health Datasets. 2017. http://www.health.alberta.ca/documents/Research-Health-Datasets.pdf

11. Karmakar C, Luo W, Tran T, Berk M, Venkatesh S. Predicting risk of suicide attempt using history of physical illnesses from electronic medical records. JMIR Mental Health. 2016; 3:3.

12. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning, with Applications in R. 6th Printing, 2015. Springer, New York.

13. Hanley JA, McNeil BJ. The meaning and use of the area under a receiving operating characteristic (ROC) curve. Radiology. 1982; 143:29-36.

14. Wang Y, Bhaskaran J, Sareen J, Bolton S, Chateau D, Bolton JM. Clinician Prediction of Future Suicide Attempts. Canadian Journal of Psychiatry. 2016; 61: 428-432.

15. Tran T, Luo W, Phung D, Harvey R, Berk M, Kennedy RL, Venkatesh S. Risk Stratification Using Data From Electronic Medical Records Better Predicts Suicide Risks Than Clinician Assessments. BMC Psychiatry. 2014; 14: 76.

16. Pisani AR, Murrie DC, Silverman MM. Risk Stratification Using Data from Electronic Medical Records Better Predicts Suicide Risks Than Clinician Assessments. Academic Psychiatry. 2016; 40: 623-629.

17. Saunders K, Brand F, Lascelles K, Hawton K. The sad truth about the SADPERSONS Scale: an evaluation of its clinical utility in self-harm patients. Emerg Med J. 2014; 31(10):796-8.

18. Katz C, Randall JR, Sareen J, et al. Predicting suicide with the SAD PERSONS scale. Depress Anxiety. 2017; 34(9):809-16.

19. Carter G, Milner A, McGill K, Pirkis J, Kapur N, Spittal MJ. Predicting suicidal behaviours

using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. Br J Psychiatry. 2017.

20. Large M, Kaneson M, Myles N, Myles H, Gunaratne P, Ryan C. Meta-analysis of longitudinal cohort studies of suicide risk assessment among psychiatric patients: heterogeneity in results and lack of improvement over time. PLoS ONE. 2016; 11(6):e0156322.

21. UNC Vision Lab. 2017.

http://image-net.org/challenges/LSVRC/2017/

22. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press. 2016. 192-195.

23. Pisani AR, Murrie DC, Silverman MM. Reformulating Suicide Risk Formulation: From Prediction to Prevention. Acad Psychiatry. 2016 Aug; 40(4):623-9

24. Hornik K. (1991). Approximation Capabilities of Multilayer Feedforward Networks. Neural Networks. 1991; 4(2):251–257.

2.10 Appendix A: Figures and Tables

Figure 1: Scatter Plot of FNN Model Configuration Performance: 8 Neurons

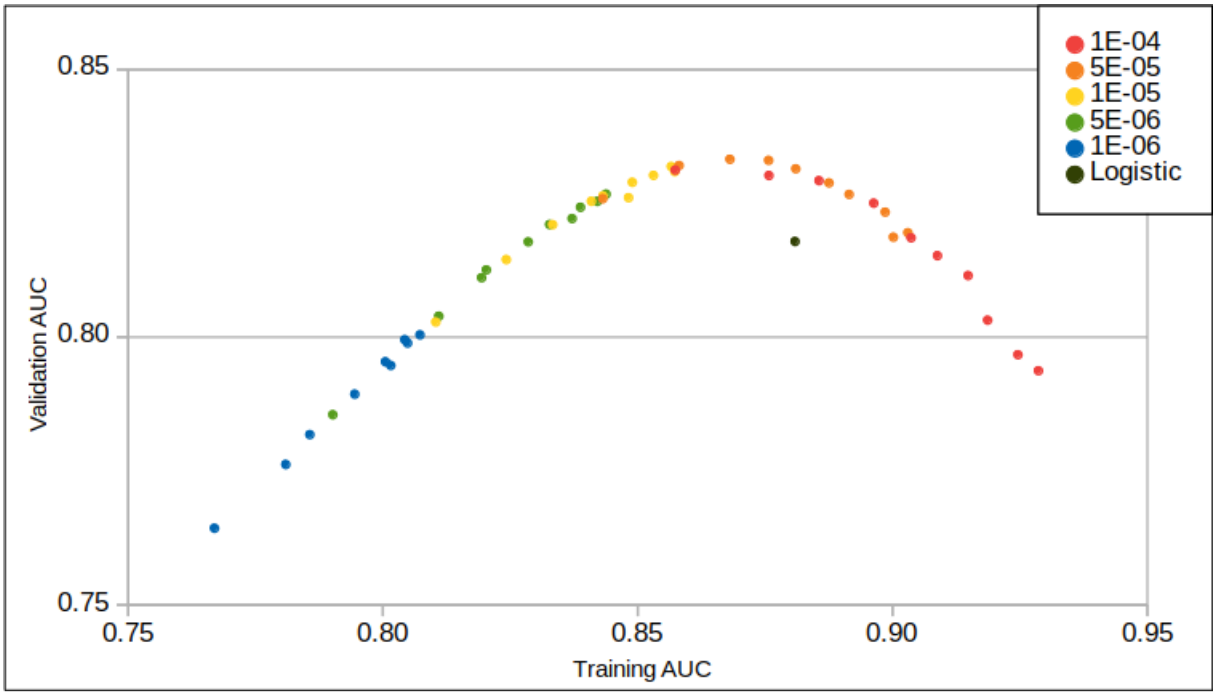


Table 1: Repeated Single Validation Set Performance Metrics, Mean and Standard Deviation

Performance Metric	**Logistic Regression 10-Fold Cross-Validation Mean**	**Feedforward Neural Network 10-Fold Cross-Validation Mean**
Area Under the Curve	0.8179	0.8352
Accuracy	0.8107	0.7998
Balanced Accuracy	0.7398	0.7547
Sensitivity	0.6531	0.6996
Specificity	0.8265	0.8098
Positive Prediction Value	0.2734	0.2691
Negative Prediction Value	0.9597	0.9642

CHAPTER 3: PREDICTING DEATH BY SUICIDE USING ADMINISTRATIVE HEALTH CARE SYSTEM DATA: CAN RECURRENT NEURAL NETWORK, ONE-DIMENSIONAL CONVOLUTIONAL NEURAL NETWORK, AND GRADIENT BOOSTED TREES MODELS IMPROVE PREDICTION PERFORMANCE?

This manuscript was published in the Journal of Affective Disorders in March, 2020. Following from the promising findings reported in the first manuscript, this manuscript was designed to evaluate the performance of cutting-edge machine learning model classes with the same modeling dataset, and to discover the time period of data required for optimal performance.

3.1 Abstract

Background

Suicide is a leading cause of death, particularly in younger persons, and this results in tremendous years of life lost.

Objective

To compare the performance of recurrent neural networks, one-dimensional convolutional neural networks, and gradient boosted trees with logistic regression and feedforward neural networks.

Methods

The modeling dataset contained 3548 persons that died by suicide and 35,480 persons that did not die by suicide between 2000 and 2016. 101 predictors were selected, and these were

assembled for each of the 40 quarters (10 years) prior to the quarter of death, resulting in 4040 predictors in total for each person. Model configurations were evaluated using 10-fold cross-validation.

Results

The optimal recurrent neural network model configuration (AUC: 0.8407), one-dimensional convolutional neural network configuration (AUC: 0.8419), and XGB model configuration (AUC: 0.8493) all outperformed logistic regression (AUC: 0.8179). In addition to superior discrimination, the optimal XGB model configuration also achieved superior calibration.

Conclusions

Although the models developed in this study showed promise, further research is needed to determine the performance limits of statistical and machine learning models that quantify suicide risk, and to develop prediction models optimized for implementation in clinical settings. It appears that the XGB model class is the most promising in terms of discrimination, calibration, and computational expense.

Limitations

Many important predictors are not available in administrative data and this likely places a limit on how well prediction models developed with administrative data can perform.

3.2 Introduction

Suicide is a leading cause of death, particularly in younger persons, and this results in tremendous years of life lost. In Alberta, Canada, suicide accounted for ten percent of the person years of life lost in persons over the age of nine between 2000 and 2017, totaling 289,078 person years of life lost [1]. During this time, suicide accounted for 23 percent of all deaths among persons 15 to 30 and 16 percent of all deaths among persons 30 to 45 in Alberta [1]. Over the same time period, the highest numbers of death by suicide in Alberta occurred in younger persons but the highest rates of death by suicide occurred in older persons [1]. Mental illness, substance misuse, parasuicide and lethality of parasuicide, suicidal ideation and intensity of suicidal ideation, social conditions and social interactions, and life events are widely recognized risk factors for suicide.

Health care service providers and health care policy providers need to able to quantify suicide risk to reduce suicide risk. Quantifying suicide risk has proven arduous [2, 3] and although statistical models have been developed that predicted suicide better than chance and better than clinicians [4, 5, 6, 7], these models have not been widely implemented, partly because the improvement in prediction performance compared to clinicians has not been striking. With the optimism surrounding artificial intelligence and machine learning, there has been discussion about whether machine learning models could improve suicide prediction [8].

In an earlier study [9] (the 'Log-FNN study'), it was shown that the feedforward neural network (FNN) class of machine learning models can improve upon logistic regression for quantifying

suicide risk with administrative health care system data in Alberta. Using a modeling dataset with 101 predictors assembled for each of the 40 quarters prior to the quarter of death (4040 predictors in total), the optimal FNN model configuration (AUC: 0.8352) outperformed logistic regression (AUC: 0.8179). The improvement in performance is promising to further explore machine learning models to quantify suicide risk.

This study will examine the performance of three machine learning model classes: Recurrent Neural Networks (RNNs), One-Dimensional Convolutional Neural Networks (1D-CNNs), and Gradient Boosted Trees (XGB). RNN and 1D-CNN models are commonly used when the order within a sequence is important, such as in natural language processing. For example, the phrases 'Scott supervises Michael' and 'Michael supervises Scott' are comprised of an identical set of words but the different ordering of the words expresses different meanings. Similarly, the order of the occurrence of predictors is important for quantifying suicide risk [12], and more recent occurrences will generally have a greater bearing on current suicide risk than less recent occurrences. XGB models were not specifically designed to model sequences but generally perform well with tabular data like the dataset in this study [13].

3.3 Neural Networks

Neural networks are a flexible class of machine learning models that were inspired by neuroscience [10]. Conceptually, a neural network model is made up of layers of neurons, with the neurons in one layer connected to the neurons in the next. Each neuron is a computational unit that multiplies its input values by a corresponding set of learnable weight parameters,

sums the multiplied values, transforms the summed value using a nonlinear activation function, and outputs the transformed value.

The first layer in a neural network model is the input layer, and each unit in the input layer contains the value of one of the predictors for a particular observation. The input layer passes all predictor values for a particular observation to each neuron in the first hidden layer. Each neuron in the first hidden layer computes a different function with the predictor values. The first hidden layer then passes its output values to each neuron in the second hidden layer, where each neuron computes a different function with its input values, and so on to the final output layer which makes a prediction.

A neural network model learns by iteratively comparing its predictions with the observed outcomes and then updating its weight parameters to improve its predictions. Neural network models have a number of hyperparameters that are set by the modeler, including the number of neurons in each hidden layer, the number of hidden layers, the learning rate, and the number of epochs. The learning rate is how much the weight parameters are adjusted at each iteration, and the number of epochs is the number of times the entire training dataset is used to update the weight parameters.

Recurrent neural networks are a class of neural networks that were designed to process sequences, and can remember or forget information from earlier steps when processing later steps in a sequence. Although FNN models are capable of representing any function [10, 11], RNN

models can learn to represent a temporal function with less parameters and less data than FNN models. This study will examine two types of RNN models: gated recurrent unit (GRU) and long short-term memory (LSTM). While similar in architecture, GRU models were developed more recently and have less parameters than LSTM models. 1D-CNN models were developed to process sequences using the Convolutional Neural Network architecture which was originally designed to process images.

3.4 Gradient Boosted Trees

Gradient boosted trees are a class of machine learning where a series of classification tree models are developed to predict the residuals of the previous model [13]. The first classification tree predicts the outcome, and then the second classification tree predicts the residuals of the predictions made by the first classification tree and so on.

XGB models have a number of hyperparameters that are set by the modeler, including the number of classification trees and the maximum classification tree depth. The number of classification trees is the number classification trees that are developed and the maximum classification tree depth is the number of times a classification model segments predictors into prediction categories.

3.5 Objective

The objective of this study is to compare the performance of RNN, 1D-CNN, and XGB models with the performance of the logistic regression and FNN models from the Log-FNN study. The

objective is not to develop optimized models for implementation but strictly to evaluate whether RNN, 1D-CNN, and XGB models are capable of providing an improvement in prediction performance compared with logistic regression and FNN models using identical modeling datasets.

It is important to explore candidate classes of models before seeking to develop models optimized for implementation because developing optimized models can be a very large undertaking. Developing optimized models using computationally expensive model classes (particularly RNNs) without reason to believe that they will outperform less computationally expensive model classes could lead to wasted time, resources, and opportunity. It is unlikely that a single prediction model could be developed and implemented everywhere, and so researchers will likely be required to develop prediction models based on the administrative health care system data available to them. This study seeks to provide direction for researchers developing prediction models by discovering the most promising prediction model class for quantifying suicide risk with administrative health care system data.

3.6 Methods

3.6.1 Data Sources

Alberta, Canada, has a population of 4.07 million people and a publicly-funded single-payer health care system with a number of administrative data systems that record the health care services of nearly its entire population [14]. A listing of the data sources and the selected predictors is available in Appendix B, but briefly, death by suicide was collected from Alberta's

vital statistics cause of death database (ICD-10 cause of death codes X60 through X84), and the predictors were collected from physician service payment claims, ambulatory care and inpatient hospitalization records, community pharmacy dispense records, and a registry containing the date Albertans qualified for a number of disease case definitions. The datasets were linked for this study using the unique Personal Health Number assigned to Albertans for the delivery of health care services. Missing predictor values occurred if a person was not a resident of Alberta during a particular quarter, and these were assigned a value of zero. A flag was included as a predictor to indicate whether a person was resident in Alberta during a particular quarter in order to distinguish missing predictor values from true zeroes.

A literature review was carried out for this study and predictors were selected from the administrative data systems if they had been shown to predict suicide or parasuicide in the literature. The predictors selected were typically related to mental health, but a number of predictors related to physical health were also selected because physical health has been shown to predict suicide [15]. Some of the predictors related to physical health may not be directly related to suicide but they were included in the modeling dataset to allow the models to learn which to regard and which to disregard.

In total, 101 predictors were selected, and these were prepared for each of the 40 quarters (10 years) prior to the quarter of death. The total number of predictors for each person was 4040 (101 predictors x 40 quarters). The modeling dataset in this study was identical to that in the

Log-FNN study but was structured with three tensor dimensions for use with RNN and 1D-CNN models (39,028 persons x 101 predictors x 40 quarters).

3.6.2 Hardware and Software

The administrative data were extracted and assembled using SAS 9.4. The analysis was performed on a desktop computer with an Ubuntu 18.04.1 LTS operating system and a GeForce GTX 1080 Ti 12GB graphics processing unit (GPU) using the NVIDIA-SMI 390.87 driver. The analysis was written in the Python programming language in a Jupyter 5.6.0 notebook in Anaconda Navigator 1.8.7. The GRU, LSTM, 1D-CNN, and FNN models were developed with Keras 2.2.2 using the TensorFlow backend with GPU support. The XGB models were developed with XGBoost 0.72 with GPU support. Keras and XGBoost are popular open-source libraries for machine learning.

3.6.3 Inclusion and Exclusion Criteria

For each person that died by suicide in Alberta between 2000 and 2016 (3548), 10 persons that did not die by suicide and were residing in Alberta in the quarter of death were randomly selected (35,480) using the proc surveyselect function in SAS 9.4. This ratio was chosen to generate a modeling dataset large enough to produce robust models while also being computationally feasible on a desktop computer with a GPU. Residents of Alberta 10 years and older were included, as 10 years is the age when suicide risk begins to manifest in the administrative data [1]. No other inclusion, exclusion, or matching criteria were applied.

As described above, 10 persons that did not die by suicide were randomly selected for each person that died by suicide, and so the outcome class distribution was imbalanced. In order to assign equal importance to both outcome classes, the models included class weights of 10 / 11 for persons that died by suicide and 1 / 11 for persons that did not die by suicide.

3.6.4 Model Configuration Evaluation

Machine learning model configurations are not evaluated with standard errors, hypothesis tests, and confidence intervals, and are instead commonly evaluated empirically with k-fold cross-validation [16]. K-fold cross-validation is a model evaluation approach that uses k validation datasets to obtain a robust estimate of expected performance with unseen data [16]. First, the modeling dataset is randomly divided into k approximately equally-sized parts (k = 5 or 10 is common). Then, a model is developed with k – 1 parts acting as a training dataset and evaluated with the remaining part acting as a validation dataset, and this process is repeated until all k parts have acted as a validation dataset once [16]. The mean of the k validation estimates is the k-fold cross-validation estimate of the expected performance of the model configuration with data the model was not developed with (unseen data) [16].

The 10-fold cross-validation receiver operating characteristic area under the curve (AUC) was chosen as the metric to evaluate model configuration performance because it has the intuitive interpretation that the AUC is the probability that the predicted risk was higher for a person that died by suicide than a person that did not [17], and because it was closely associated with sensitivity, specificity, positive prediction value (PPV), and negative prediction value (NPV).

The compute time required for the Log-FNN study was approximately 3 days. RNN models are generally more computationally expensive than FNN models, and it was estimated that evaluating LSTM and GRU model configurations with the same range of neuron and learning rate settings as the FNN models in the Log-FNN study would require approximately 50 days of compute time. To reduce the compute time required to find the optimal GRU and LSTM model configurations, a single neuron configuration (8 neurons) was chosen for the RNN models. A single neuron configuration was considered sufficient for evaluation in this study rather than the five (8, 16, 32, 64, 128) in the Log-FNN study because all of the neuron configurations in the Log-FNN study achieved essentially identical optimal 10-fold cross-validation AUCs and it was expected that this would be the case in this study as well.

To reduce the compute time further, the RNN model evaluation occurred in two stages. In the first stage, GRU and LSTM model configurations were evaluated with 8 neurons, learning rates of 1e-4, 5e-5, 1e-5, 5e-6, and 1e-6, and a sparse range of 275, 500, 750, and 1000 epochs. In the second stage, GRU and LSTM model configurations were evaluated with 8 neurons, the most promising learning rate from stage 1, and a more refined range of 50 to 1000 epochs in increments of 50 epochs. The most promising learning rate for the GRU model configurations was 1e-4 and the most promising learning rate for the LSTM model configurations was 5e-5.

The 1D-CNN hyperparameters evaluated in this study were the one-dimensional convolutional kernel size (1, 2, 4, 6, 8), the learning rate (1e-4, 5e-5, 1e-5, 5e-6, 1e-6), and the number of epochs (50 to 1000 in increments of 50). The number of filters is also a hyperparameter but

after preliminary exploration with a range of filters, it was decided to default to 8 filters. The XGB hyperparameters evaluated in this study were the number of classification trees (50 to 1000 in increments of 50) and the maximum classification tree depth (1, 2, 3, 4, 5). The learning rate is also a hyperparameter but after preliminary exploration with a range of learning rates, it was decided to use the default setting in the XGBoost software.

The RNN model configuration evaluation described above took approximately 7 days of compute time, compared to the estimate of over 50 days for evaluation with the same range of neuron and learning rate settings from the Log-FNN study. The 1-D CNN model configuration evaluation took around 5 days of compute time and the XGB model configuration evaluation took around 7 hours of compute time.

3.6.5 Smoothed Performance Trajectories

The objective of evaluating model configurations is to discover the configuration with the best expected performance with unseen data. Although 10-fold cross-validation provides a robust estimate of the expected performance of a model configuration with unseen data, the estimate is unlikely to be exactly equal to the true expected performance of that model configuration with unseen data. Quadratic polynomial lines will be used in this study when evaluating neural network model configurations to smooth out the variability in the 10-fold cross-validation AUC estimates over the range of epochs.

Smoothed performance trajectories provide a better sense of the expected performance of a model configuration with unseen data and provide a cleaner visual depiction of the performance trajectories. As an illustration, Figure 1 shows the performance trajectories of the FNN model configurations with 8 neurons from the Log-FNN study as a scatter plot of the 10-fold cross-validation training AUC versus the 10-fold cross-validation validation AUC over different epoch settings (50, 100, 150, 200, 250, 300, 350, 400, 450, 500). The 10-fold cross-validation estimate for logistic regression is represented by a single point because there was only a single model configuration.

3.7 Results

3.7.1 Discrimination

The 10-fold cross-validation AUC estimates were 0.8407 for the optimal GRU model configuration, 0.8356 for the optimal LSTM model configuration, 0.8419 for the optimal 1D-CNN model configuration, and 0.8493 for the optimal XGB model configuration. In addition to the AUC, a number of other 10-fold cross-validation performance metrics were computed and are included in Table 1. The optimal GRU model configuration performed slightly better than the optimal LSTM model configuration on every performance metric. The optimal GRU model configuration also performed slightly better than the optimal FNN model configuration from the Log-FNN study on every performance metric. The optimal neural network models had higher sensitivity, while the optimal XGB model configuration had slightly lower sensitivity with higher specificity and PPV.

The optimal GRU model configuration had greater sensitivity (0.7130 vs 0.6531) than the logistic regression model from the Log-FNN study, with similar specificity (0.8097 vs 0.8265), PPV (0.2728 vs 0.2734), and NPV (0.9658 vs 0.9597). The optimal 1D-CNN model configuration had greater sensitivity (0.7207 vs 0.6531) than the logistic regression model from the Log-FNN study, with similar PPV (0.2721 vs 0.2734) and NPV (0.9666 vs 0.9597) and lower specificity (0.8066 vs 0.8265). The optimal XGB model configuration had greater sensitivity (0.6983 vs 0.6531) and PPV (0.2901 vs 0.2734) than the logistic regression model from the Log-FNN study, with similar specificity (0.8290 vs 0.8265) and NPV (0.9648 vs 0.9597).

3.7.2 Calibration

The calibration of logistic regression and the optimal XGB model configuration was compared using calibration curves. Calibration curves compare the predicted probability of the outcome with the actual probability of the outcome in the modeling dataset. This study used a case-control study design and so the probabilities of the outcome in the modeling dataset are not representative of a realistic setting; however, calibration curves were compared to determine which model class can generally be expected to achieve better calibration.

The logistic regression and XGB models included class weights (10 / 11 for persons that died by suicide and 1 / 11 for persons that did not die by suicide) so that equal importance was assigned both outcome classes, and so the predicted probabilities were calibrated as though the modeling dataset contained balanced outcome classes. To evaluate the models calibrated

to the actual risk of death by suicide in the modeling dataset, Platt calibration was used [18]. Platt calibration uses logistic regression to calibrate predicted probabilities into actual probabilities.

The modeling dataset was randomly divided into training (80 percent) and validation (20 percent) datasets. The logistic regression and XGB models were developed with the training dataset and validated with the validation dataset. The XGB model with Platt calibration achieved better calibration than the logistic regression model with Platt calibration for both the training and validation datasets (Figures 2 and 3), particularly for predicted probabilities higher than 0.2. Both models tended to produce higher predicted probabilities for higher actual probabilities but the calibration curves for the XGB model were far less variable.

3.7.3 Most Recent Quarters

To examine temporality from another perspective, the optimal 8-neuron FNN model configuration from the Log-FNN study (learning rate of 5e-5), the optimal 8-neuron GRU configuration (learning rate of 1e-4), the optimal 1D-CNN configuration (kernel size of 2, learning rate of 5e-5), and all XGB model configurations were compared using modeling datasets containing the most recent 2, 4, 8, 12, and 16 quarters rather than all 40 quarters.

The FNN model configuration achieved optimal performance using the most recent 4 quarters (AUC: 0.8406) which was higher than the optimal FNN performance with all 40 quarters (AUC: 0.8352). The GRU model configuration achieved optimal performance using the most recent 16 quarters (AUC: 0.8415) which was similar to the optimal GRU performance with all 40 quarters

(AUC: 0.8407). The 1D-CNN model configuration achieved optimal performance using the most recent 12 quarters (AUC: 0.8415) which was similar to the optimal 1D-CNN performance with all 40 quarters (AUC: 0.8419). The XGB model configuration achieved optimal performance using the most recent 4 quarters (AUC: 0.8500) which was similar to the optimal XGB performance with all 40 quarters (AUC: 0.8493).

Examining the smoothed performance trajectories of the FNN and GRU models (see: Figures 4 and 5), the FNN model configuration using the most recent 4 quarters, the GRU model configuration using the most recent 16 quarters had the highest smoothed AUCs. Performance increased with more quarters until a maximum was reached, after which additional quarters resulted in slowly decreasing performance. This was also the case with the 1D-CNN and XGB model configurations but these figures are not included for the sake of brevity.

3.8 Discussion

The objective of this study is to compare the performance of RNN, 1D-CNN, and XGB model configurations with the performance of the logistic regression and FNN model configurations from the Log-FNN study. Although the optimal GRU (AUC: 0.8407) and the optimal 1D-CNN (AUC: 0.8419) model configurations achieved better discrimination than the optimal FNN model configuration from the Log-FNN study (AUC: 0.8354) using the analytic dataset with all 40 quarters, the improvement in performance was slight. The smoothed performance trajectories of the optimal model configurations using analytic datasets with 2, 4, 8, 12, and 16 quarters showed that the optimal GRU (16 quarters) and optimal 1D-CNN (12 quarters) model

configurations outperformed the optimal FNN model configuration (4 quarters) but again the improvement in performance was slight, while the optimal XGB model configuration (4 quarters) outperformed all of the neural network models (see: Figure 6). In addition to superior discrimination, the optimal XGB model achieved superior calibration compared with logistic regression.

The XGB model class was by far the least computationally expensive and predicted death by suicide better than the neural network model classes in terms of discrimination and calibration. It appears from this study and from the Log-FNN study that XGB models are promising for future research on quantifying suicide risk but that FNN, RNN, and 1D-CNN models do not justify their large computational expense and longer temporal data requirements.

An interesting finding from this study is that using analytic datasets with increasing quarters eventually led to slowly decreasing performance. Performance increased with more quarters until a maximum was reached (see: Figures 4 and 5), after which additional quarters resulted in decreasing performance. It is possible that less recent data has no prediction utility and only increases noise, or it is possible that there were not enough persons in the analytic dataset to allow models to learn functions that made full use of less recent data.

Also interesting is that the optimal FNN model configuration which used the most recent 4 quarters achieved performance close to the optimal GRU and optimal 1D-CNN model configurations which used more quarters, and all were outperformed by the optimal XGB

model configuration which used the most recent 4 quarters (see: Figure 6). The prevention framework suggested by Pisani et al. [12] considers suicide risk to have two components: risk status (risk relative to other persons) and risk state (risk relative to prior personal states). Although not definitive, the results suggest that risk state over the past year is most important for quantifying suicide risk and that considering risk states over longer time periods will not result in improvements in quantifying suicide risk.

Further research is needed to determine whether prediction models can be developed that will be attractive to health care service providers and health care policy providers. Statistical prediction models have been developed that outperform clinicians when predicting the risk of suicidality [19, 20], but these models have not been widely adopted in clinical settings. Instead, risk scales are commonly used in clinical settings but risk scales have limited utility for quantifying suicidality risk [21, 22, 23, 24, 25]. Although this study did not seek to develop a prediction model for clinical practice, the ultimate goal of this research is to take the first steps toward the development of prediction models that have optimal prediction performance and optimal relevance for health care service providers and health care policy providers.

In order to develop prediction models that will be attractive to health care service providers and health care policy providers, further research is needed. For example, it is unlikely that all 101 predictors in the modeling dataset are required for optimal or near-optimal performance. Reducing the number of predictors would simplify the prediction models and also reduce the burden of data collection. A smaller number of predictors would also help to understand which

predictors are most important for quantifying suicide risk. In addition, this study quantified suicide risk within 90 days but it would also be valuable for further research to evaluate how far into the future suicide risk can be reliably quantified. Further research is also necessary to achieve consensus on the preferred performance characteristics (preferred values of sensitivity, specificity, PPV, NPV) for prediction models that quantify suicide risk.

3.9 Limitations

There were three primary limitations in this study: the case-control sampling design, the volume of data, and the inherent limitations of administrative data. The case-control sampling design and data volume limitations arose because of computational considerations and could be addressed by future research, but the inherent limitations of administrative data cannot be overcome as easily.

First, a case-control sampling design was used to generate a modeling dataset that was computationally feasible on a desktop computer with a GPU. The case-control sampling design is useful for comparing relative model discrimination and calibration but the actual probabilities are not meaningful. Suicide is a rare event, and a modeling dataset that contained enough persons that died by suicide to develop robust prediction models and that also had a realistic risk of death by suicide might need to contain hundreds of thousands or even millions of persons depending on the setting.

Second, the modeling dataset in this study contained a large volume of data compared to many studies of suicide but it may not be a large enough volume for FNN, GRU, 1D-CNN, and XGB models to learn a more complex function than the logistic regression function, even with a 1:10 case-control sampling design. This may explain why the FNN, GRU, 1D-CNN, and XGB model configurations outperformed the logistic regression model by a smaller margin than might have been hoped for considering the optimism surrounding machine learning and artificial intelligence. To develop prediction models with very large datasets, researchers may require virtual server services such as Amazon Web Services EC2 or Google Cloud AI Platform. This study was not able to use virtual server services due to legislative restrictions.

Third, administrative data have inherent limitations. The predictors available in the administrative data were not collected for the purposes of quantifying suicide risk and many important predictors were not available. This is likely the most fundamental limitation of administrative data for quantifying suicide risk but this limitation could diminish if electronic health care system data becomes richer and the ability to link with non-health care system data improves. Another limitation of administrative data is that temporal precision is critical with suicide because risk can escalate to crisis in a very short period of time and administrative data may not be refined enough or timely enough to predict crisis states. That said, health care service providers and health care policy providers may prefer to manage suicide risk before it reaches a crisis, and administrative data might be the best source of data for quantifying non-crisis suicide risk.

3.10 References

1. Alberta Vital Statistics. Cause of Death database; ICD-10: X60 through X84. 2019.

2. Mulder R, Newton-Howes G, Coid JW. The futility of risk prediction in psychiatry. Br J Psychiatry 2016; 209:271-272.

3. Large M, Kaneson M, Myles N, Myles H, Gunaratne P, Ryan C. Meta-analysis of longitudinal cohort studies of suicide risk assessment among psychiatric patients: heterogeneity in results and lack of improvement over time. PLoS ONE. 2016; 11(6):e0156322.

4. Chan MK, Bhatti H, Meader N, Stockton S, Evans J, O'Connor RC, et al. Predicting suicide following self-harm: systematic review of risk factors and risk scales. Br J Psychiatry. 2016; 209(4):277-83.

5. Huang X, Ribiero JD, Musacchio KM, Franklin JC. Demographics as predictors of suicidal thoughts and behaviors: A meta-analysis. PLoS ONE. 2017; 12(7):e0180793.

6. Carter G, Milner A, McGill K, Pirkis J, Kapur N, Spittal MJ. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. Br J Psychiatry. 2017.

7. Ribeiro JD, Franklin JC, Fox KR, Bentley KH, Kleiman EM, Chang BP, Nock MK. Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. Psychological Medicine. 2016; 46:225–236.

8. Ribeiro JD, Franklin JC, Fox KR, Bentley KH, Kleiman EM, Chang BP, Nock MK. Suicide as a complex classification problem: machine learning and related techniques can advance suicide prediction – a reply to Roaldset. Psychological Medicine. 2016; 46:2009–2010.

9. Sanderson M, Bulloch A, Wang J, Williamson T, Patten S. Predicting Death by Suicide Using Administrative Health Care System Data: Can Feedforward Neural Network Models Improve Upon Logistic Regression Models? Journal of Affective Disorders. 2019; 257:741-747

10. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press. 2016. 192-195.

11. Hornik K. (1991). Approximation Capabilities of Multilayer Feedforward Networks. Neural Networks. 1991; 4(2):251–257.

12. Pisani AR, Murrie DC, Silverman MM. Reformulating Suicide Risk Formulation: From Prediction to Prevention. Acad Psychiatry. 2016 Aug; 40(4):623-9

13. The XGBoost Contributors. https://xgboost.readthedocs.io/en/latest/tutorials/model.html

14. Alberta Health: Overview of Administrative Health Datasets. 2017. http://www.health.alberta.ca/documents/Research-Health-Datasets.pdf

15. Karmakar C, Luo W, Tran T, Berk M, Venkatesh S. Predicting risk of suicide attempt using history of physical illnesses from electronic medical records. JMIR Mental Health. 2016; 3:3.

16. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning, with Applications in R. 6th Printing, 2015. Springer, New York.

17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiving operating characteristic (ROC) curve. Radiology. 1982; 143:29-36.

18. https://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html

19.Tran T, Luo W, Phung D, Harvey R, Berk M, Kennedy RL, Venkatesh S. Risk Stratification Using Data From Electronic Medical Records Better Predicts Suicide Risks Than Clinician Assessments. BMC Psychiatry. 2014; 14: 76.

20. Pisani AR, Murrie DC, Silverman MM. Risk Stratification Using Data from Electronic Medical Records Better Predicts Suicide Risks Than Clinician Assessments. Academic Psychiatry. 2016; 40: 623-629.

21. Saunders K, Brand F, Lascelles K, Hawton K. The sad truth about the SADPERSONS Scale: an evaluation of its clinical utility in self-harm patients. Emerg Med J. 2014; 31(10):796-8.

22. Katz C, Randall JR, Sareen J, et al. Predicting suicide with the SAD PERSONS scale. Depress Anxiety. 2017; 34(9):809-16.

23. Chan MK, Bhatti H, Meader N, Stockton S, Evans J, O'Connor RC, et al. Predicting suicide following self-harm: systematic review of risk factors and risk scales. Br J Psychiatry. 2016; 209(4):277-83.

24. Carter G, Milner A, McGill K, Pirkis J, Kapur N, Spittal MJ. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. Br J Psychiatry. 2017.

25. Large M, Kaneson M, Myles N, Myles H, Gunaratne P, Ryan C. Meta-analysis of longitudinal cohort studies of suicide risk assessment among psychiatric patients: heterogeneity in results and lack of improvement over time. PLoS ONE. 2016; 11(6):e0156322.

3.11 Appendix A: Figures and Tables

Figure 1: Smoothed Performance Trajectories for the 8 Neuron FNN Model Configurations and Logistic Regression from the Log-FNN Study

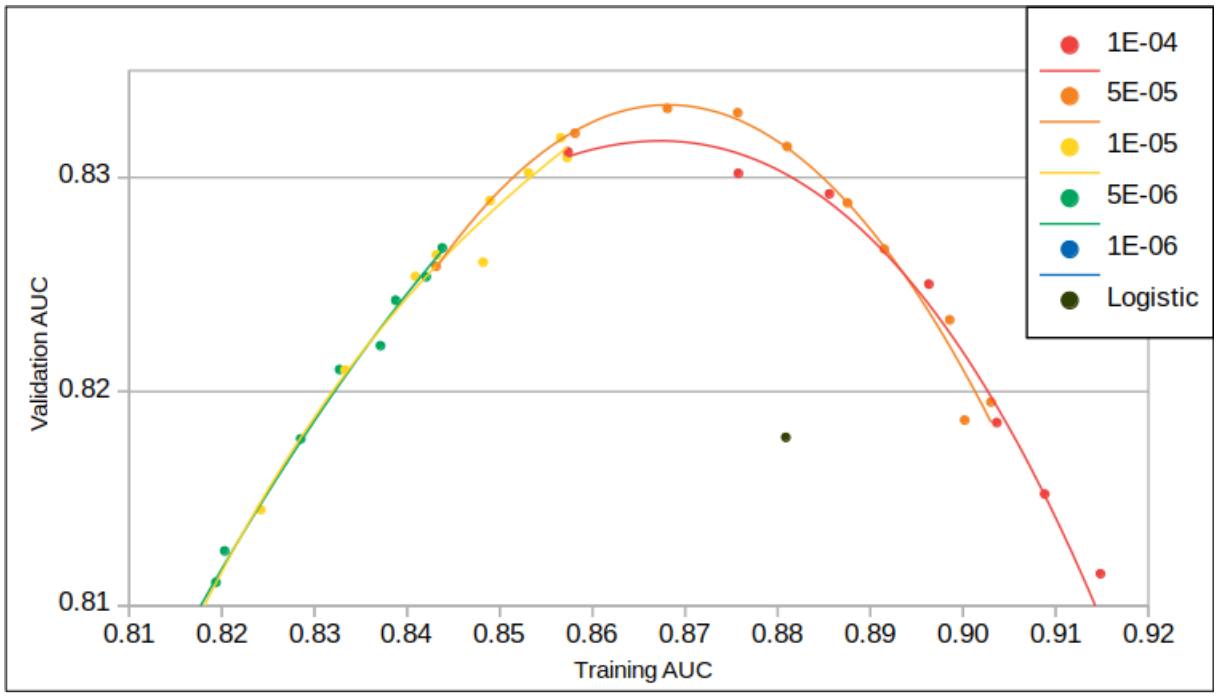


Figure 2: Calibration Curves, Logistic Regression, Platt Calibration

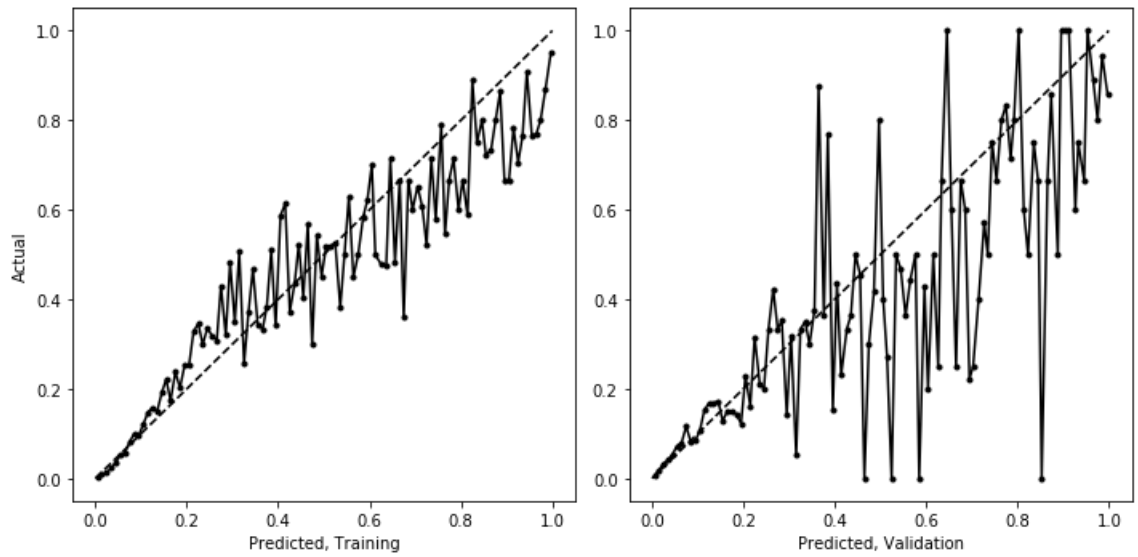


Figure 3: Calibration Curves, Gradient Boosted Trees, Platt Calibration

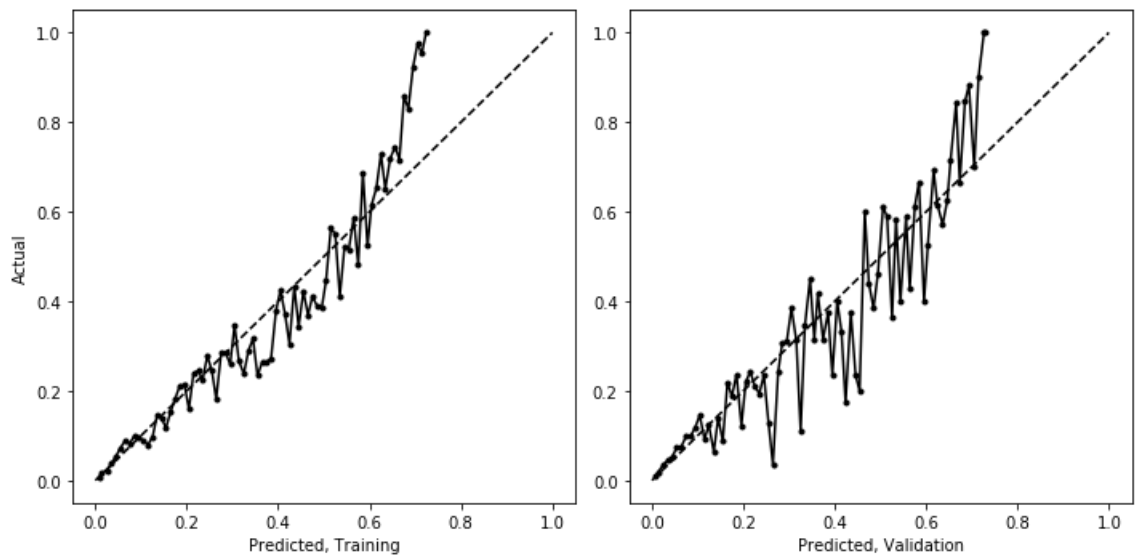


Figure 4: Smoothed Performance Trajectories for the FNN Model Configurations, Quarters

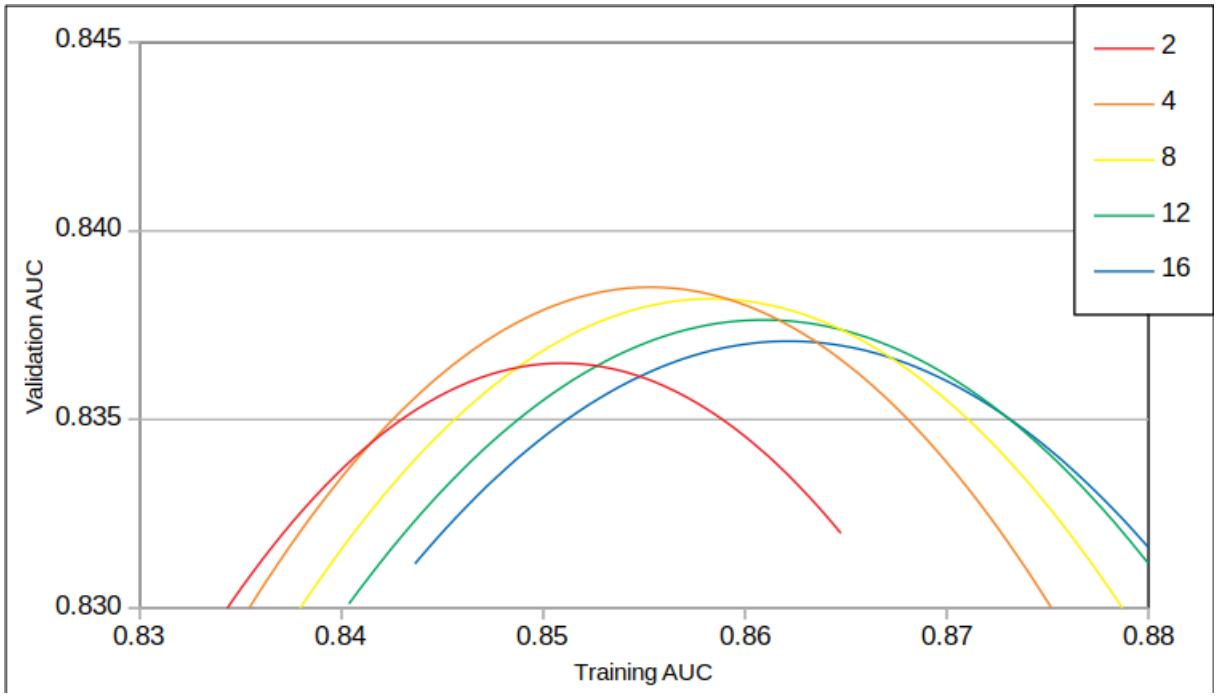


Figure 5: Smoothed Performance Trajectories for the GRU Model Configurations, Quarters

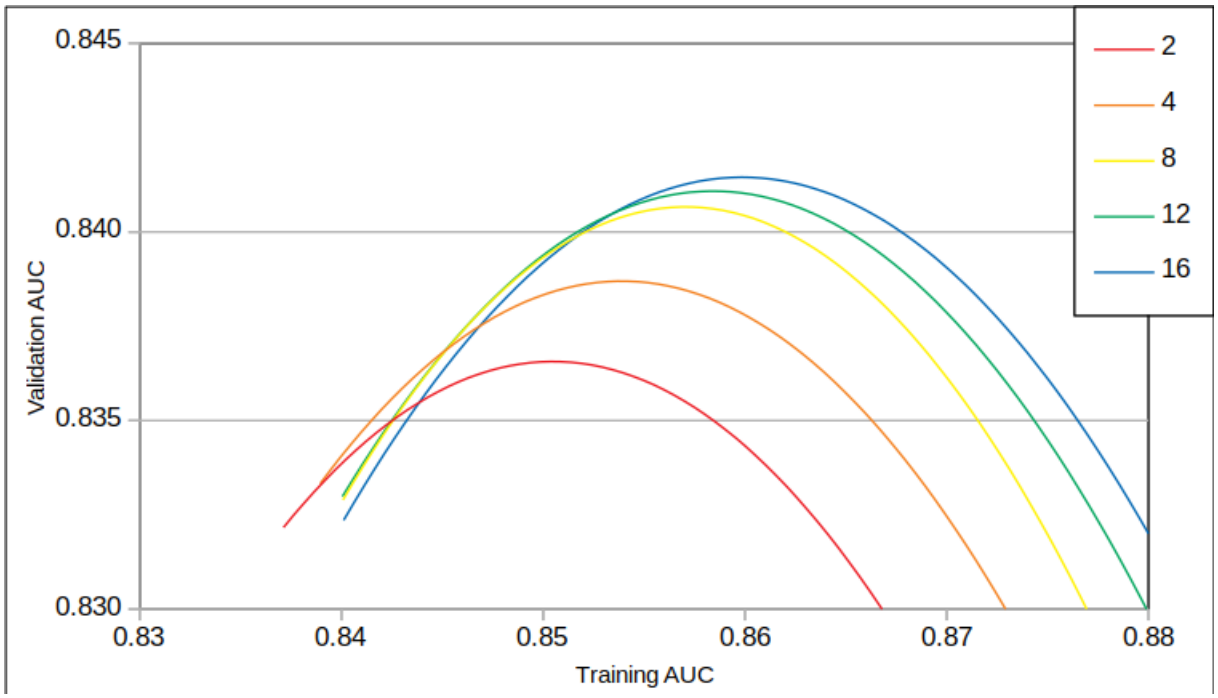


Figure 6: Performance Trajectories for the Optimal FNN, GRU, CNN, and XGB Model Configurations, Quarters

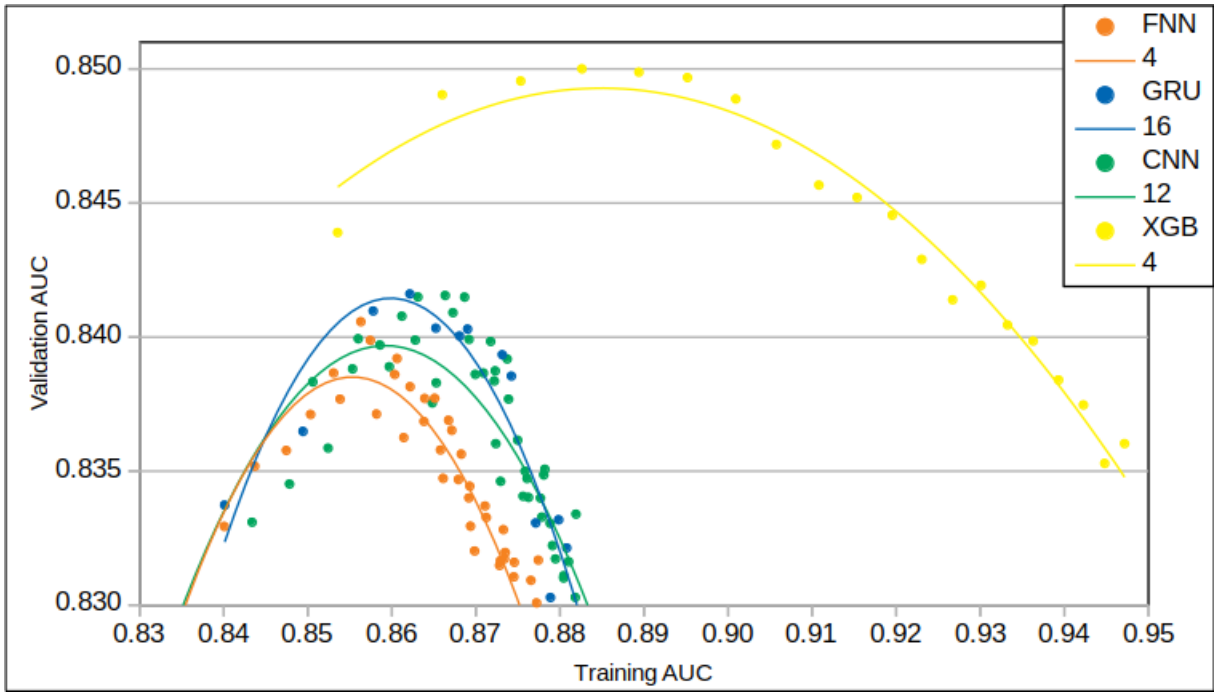


Table 1: 10-Fold Cross-Validation Performance Metrics, Mean

Performance Metric	LSTM Mean	GRU Mean	1D-CNN Mean	XGB Mean
Area Under the Curve	0.8356	0.8407	0.8419	0.8493
Accuracy	0.7947	0.8009	0.7988	0.8171
Balanced Accuracy	0.7550	0.7614	0.7637	0.7637
Sensitivity	0.7066	0.7130	0.7207	0.6983
Specificity	0.8035	0.8097	0.8066	0.8290
Positive Prediction Value	0.2647	0.2728	0.2721	0.2901
Negative Prediction Value	0.9648	0.9658	0.9666	0.9648

CHAPTER 4: PREDICTING DEATH BY SUICIDE FOLLOWING AN EMERGENCY DEPARTMENT VISIT FOR PARASUICIDE WITH ADMINISTRATIVE HEALTH CARE SYSTEM DATA AND GRADIENT BOOSTED TREES

This manuscript was published in EClinicalMedicine in April, 2020. This manuscript was designed to apply the findings from the first two manuscripts to a clinical setting as the next step towards evaluating whether prediction models developed with administrative health care system data can achieve promising performance in a clinical setting.

The first two manuscripts found that the gradient boosted trees model class achieved the best performance and was also the least computationally expensive, and that 8 quarters of temporal data or less was required for optimal gradient boosted trees performance. These were the analytic foundations of this manuscript.

4.1 Abstract

Suicide is a leading cause of death worldwide and results in a large number of person years of life lost. There is an opportunity to evaluate whether administrative health care system data and machine learning can quantify suicide risk in a clinical setting.

The objective was to compare the performance of prediction models that quantify the risk of death by suicide within 90 days of an ED visit for parasuicide with predictors available in administrative health care system data.

The modeling dataset was assembled from 5 administrative health care data systems. The data systems contained nearly all of the physician visits, ambulatory care visits, inpatient hospitalizations, and community pharmacy dispenses, of nearly the entire 4.07 million persons in Alberta, Canada. 101 predictors were selected, and these were assembled for each of the 8 quarters (2 years) prior to the quarter of death, resulting in 808 predictors in total for each person. Prediction model performance was validated with 10-fold cross-validation.

The optimal prediction model achieved promising discrimination (AUC: 0.88) and calibration that could lead to clinical applications. The 5 most important predictors in the optimal gradient boosted trees model each came from a different administrative health care data system.

The combination of predictors from multiple administrative data systems and the combination of personal and ecologic predictors resulted in promising prediction performance. Further research is needed to develop prediction models optimized for implementation in clinical settings.

4.2 Introduction

Although death by suicide is a rare event, it is an important cause of death because most deaths by suicide are premature deaths and result in a large number of years of life lost. In the Canadian province of Alberta, between 2000 and 2018, the suicide rate was 14 per 100,000 person-years, and 96 percent of deaths by suicide occurred in persons younger than 75 resulting

in 290,490 years of life lost [1]. 84 percent of deaths by suicide occurred in persons younger than 60 and 53 percent occurred in persons younger than 45 [1].

There are a number of risk factors that are widely recognized for death by suicide, including mental illness, substance misuse, parasuicide and lethality of parasuicide, suicidal ideation and intensity of suicidal ideation, social conditions and social interactions, and life events. Although many risk factors for suicide are known, quantifying suicide risk is difficult [2, 3, 4] and this makes suicide prevention a challenge for health care service providers and health care policy providers. Risk scales are often used in clinical settings but it has been shown that risk scales have limited utility for quantifying suicidality risk [5, 6, 7, 8, 9]. Statistical models have been developed to quantify suicidality risk but these models have not been widely implemented, even though the models outperformed clinicians when compared [10, 11]. In Canada, large amounts of data are collected during the administration of the health care system. This data provides an opportunity to explore whether quantifying suicide risk with machine learning models using administrative data can achieve performance that is potentially capable of guiding preventive actions.

In earlier studies [12, 13], it was found that the feedforward neural network, recurrent neural network, one-dimensional convolutional neural network, and gradient boosted trees classes of machine learning models can improve upon logistic regression when quantifying suicide risk with administrative health care system data in Alberta. The optimal feedforward neural network (AUC: 0.8352), recurrent neural network (AUC: 0.8407), one-dimensional convolutional neural network (0.8419), and gradient boosted trees (AUC: 0.8493) model configurations

outperformed logistic regression (AUC: 0.8179). It was found that gradient boosted trees model configurations outperformed the neural network model configurations and required far less computational resources.

Further, although 10 years (40 quarters) of temporal data was available in the modeling dataset, and recurrent neural networks and one-dimensional convolutional neural networks are designed to process sequences, the optimal recurrent neural network and one-dimensional convolutional neural network model configurations did not materially outperform the optimal feedforward neural network model configuration, required more data to achieve optimal performance, and were far more computationally expensive. The optimal gradient boosted trees and feedforward neural network model configurations required less than two years of temporal data for optimal performance.

While the earlier studies were designed to identify the most promising model classes and the temporal period required to achieve optimal performance, they used a case-control study design in order to include as many instances of death by suicide as possible in the modeling dataset. The resulting modeling dataset was not representative of a health care setting where a predictive model may have clinical utility. This study seeks to extend the findings of the earlier studies to a realistic health care setting: emergency department (ED) visits for parasuicide (self-harm that did not result in death, regardless of intent). ED visits for parasuicide present a unique opportunity for suicide prevention because these visits identify persons with a high risk of death by suicide (1 in 125 in this study compared with the overall Alberta 90-day risk of 1 in

29,000) and provide opportunities to reduce the imminent risk of suicide and to establish continuity of care to reduce suicide risk following discharge [14]. If the risk of death by suicide following an ED visit for parasuicide could be quantified, then health care service providers and health care policy providers may be able to better target prevention efforts. For example, inpatient admission can be used as a preventive action, but a number of other treatment options are available (discharge with routine follow-up, discharge with urgent follow-up, assertive outreach, etc.), and being able to quantify suicide risk would help health care service providers decide on the best treatment option.

The objective of this study is to compare the performance of logistic regression and gradient boosted trees (XGB) models for quantifying the risk of death by suicide within 90 days of an ED visit for parasuicide with predictors available in administrative health care system data.

4.3 Objective

The objective of this study is to compare the performance of logistic regression and gradient boosted trees (XGB) models for quantifying the risk of death by suicide within 90 days of an ED visit for parasuicide (self-harm that did not result in death, regardless of intent) with predictors available in administrative health care system data.

4.4 Methods

4.4.1 Data Sources

A literature review was carried out for this study. The goal of the literature review was to identify predictors that have been used to predict suicide or parasuicide. The majority of predictors were identified from clinical assessment tools and statistical prediction models. Predictors were selected from administrative data systems if they had been shown to predict suicide or parasuicide in the literature review. A complete listing of the administrative data sources and the selected predictors is available in Appendix B. The data sources contain nearly all of the physician visits, ambulatory care visits, inpatient hospitalizations, and community pharmacy dispenses, of nearly the entire 4.07 million persons in Alberta, Canada [15]. Death by suicide was collected from the vital statistics cause of death database (ICD-10 cause of death codes X60 through X84), and the predictors were collected from physician service payment claims, ambulatory care and inpatient hospitalization records, community pharmacy dispense records, and a registry containing the date of qualification for a number of disease case definitions. The data were linked using the unique Personal Health Number assigned to Albertans for the delivery of health care services.

Parasuicide was defined as an ED visit for self-harm that did not result in death, regardless of intent. The term 'parasuicide' is used rather than the term 'attempted suicide' because intent cannot be determined with the administrative data used in this study. ED visits coded with a disposition of "death on arrival (DOA): patient is dead on arrival to the ambulatory care service" and "death after arrival (DAA): patient expires after initiation of the ambulatory care visit" were

excluded because these were considered deaths by suicide. Persons with a date of death in the vital statistics data on the same day as the most recent ED visit for parasuicide – whatever the cause of death – were also excluded because there would be no opportunity for follow-up and would not be relevant to decisions made by clinicians in the ED.

All persons with an ED visit for parasuicide between 2010 and 2017 were extracted from the ambulatory care data system. The most recent ED visit for parasuicide was selected, and the predictors were assembled for each of the most recent 8 quarters because our earlier work [12, 13] showed that only the most recent 8 quarters were required for optimal prediction performance. In total, 101 predictors were selected, and these were prepared for each of the 8 quarters prior to the most recent ED visit for parasuicide. The modeling dataset did not include any information following the most recent ED visit for parasuicide. The predictors selected were primarily related to mental health, but predictors related to physical health were also selected because physical health has been shown to predict suicide [16]. The predictors related to physical health may not be directly related to suicide but they were included in the modeling dataset to allow the models to learn which (if any) contribute to quantifying suicide risk. The total number of predictors for each person was 808 (101 predictors x 8 quarters). The outcome was death by suicide within 90 days of the most recent ED visit for parasuicide.

There were 268 persons that died by suicide within 90 days and 33,426 persons that did not, and so the outcome class distribution was imbalanced. In order to assign equal importance to

both outcome classes, the models included class weights of 124 / 125 for persons that died by suicide and 1 / 125 for persons that did not die by suicide.

4.4.2 Hardware and Software

The administrative data were extracted and assembled using SAS 9.4. The analysis was performed on a desktop computer with an Ubuntu 18.04.1 LTS operating system and a GeForce GTX 1080 Ti 12GB graphics processing unit (GPU) using the NVIDIA-SMI 390.87 driver. The analysis was written in the Python programming language in a Jupyter 5.6.0 notebook in Anaconda Navigator 1.8.7. The logistic regression models and calibration curves were developed using scikit-learn 0.20.0 [17]. The XGB models were developed with XGBoost 0.72 [18] with GPU support.

4.4.3 Model Configuration Evaluation

K-fold cross-validation is a model evaluation approach that uses k validation datasets to obtain a robust estimate of expected performance with unseen data [19]. The 10-fold cross-validation area under the receiver operating characteristic curve (AUC) was chosen as the metric to evaluate model configuration performance because it has the intuitive interpretation that the AUC is the probability that the predicted risk was higher for a person that died by suicide than a person that did not [20], and because it was closely associated with sensitivity, specificity, positive prediction value (PPV), and negative prediction value (NPV).

The logistic regression and XGB model configurations were evaluated with the most recent 1, 2, 4, 6, and 8, quarters of data. The scikit-learn library used to develop the logistic regression models applies a L2 regularization penalty (often called 'ridge regression') by default [21]. The L2 regularization penalty adds the sum of the squared beta parameters to the loss function that the logistic regression model seeks to minimize. This has the effect of penalizing large beta parameter values and can help prevent overfitting. To evaluate the logistic regression model configurations without a regularization penalty (the default in most statistical software), the C parameter in scikit-learn was assigned a value 1,000,000. The C parameter is the inverse of regularization strength, and so a regularization strength of 1 / 1,000,000 essentially disables regularization. The XGB hyperparameters evaluated in this study were the number of classification trees (10 to 200 in increments of 10) and the maximum classification tree depth (1, 2, 3, 4, 5). The learning rate and gamma are also XGB hyperparameters but after preliminary exploration with a range of settings, it was decided to use the default settings in the XGBoost library (gamma = 0, learning rate = 0.1) because adjusting the default settings did not result in performance improvements.

4.4.4 Role of Funding

There was no funding for this study.

4.5 Results

4.5.1 Discrimination

The 10-fold cross-validation AUC estimates for logistic regression with the L2 regularization penalty disabled (C parameter = 1 / 1,000,000) using the most recent 1, 2, 4, 6, and 8, quarters were 0.8113, 0.7760, 0.7361, 0.6988, 0.6758, respectively. Logistic regression with the L2 regularization penalty disabled was overfit to the training data, and the overfitting was more severe with additional quarters of temporal data. Conversely, the 10-fold cross-validation AUC estimates for logistic regression with the L2 regularization penalty enabled (the default in the scikit-learn library) using 1, 2, 4, 6, and 8, quarters were 0.8590, 0.8632, 0.8572, 0.8454, 0.8392, respectively.

The 10-fold cross-validation AUC estimate was 0.8786 for the optimal XGB model configuration (2 quarters of data, 70 classification trees, maximum tree depth of 2). The performance of the XGB model configurations with the most recent 2, 4, 6, and 8, quarters was essentially indistinguishable but the XGB model configurations using 2 and 4 quarters tended to have slightly higher optimal AUC estimates.

In addition to the AUC, a number of other 10-fold cross-validation performance metrics were computed and are included in Table 1. The optimal XGB model configuration performed better than the optimal logistic regression model configuration with L2 regularization disabled on every performance metric. The optimal XGB model configuration had a higher sensitivity

(0.8912 vs 0.8420) than the optimal logistic regression model configuration with L2 regularization enabled but a lower specificity (0.6876 vs 0.7429).

4.5.2 Calibration

The calibration of prediction models is often evaluated by comparing predicted probabilities with actual probabilities, commonly called a 'calibration curve'. The calibration of the optimal logistic regression (2 quarters of data, L2 regularization) and XGB (2 quarters of data, 70 classification trees, maximum tree depth of 2) model configurations from above were evaluated using calibration curves. To evaluate calibration with unseen data, the modeling dataset was divided into a training dataset (80 percent) and a validation dataset (20 percent). With modeling datasets that have a small number of instances of the outcome, random divisions of the modeling dataset into training and validation datasets can sometimes result in a validation dataset with a disproportionate number of instances of the outcome which can result in poor calibration. Stratified random sampling based on the outcome is commonly used to ensure that the validation dataset has a proportionate number of instances of the outcome. The division of the modeling dataset into training and validation datasets was stratified based on the outcome to ensure that both datasets had the same proportion of deaths by suicide as the modeling dataset. The models were developed with the training dataset and evaluated with the validation dataset.

Figures 1 and 2 show the calibration curves for the logistic regression and XGB models, evaluated on both the training and validation datasets. The predicted probability generally

increased as the actual probability increased but the agreement between the predicted and actual probabilities was variable. The variability was mainly due to the small number of instances of death in the modeling dataset. For example, if the XGB model predicted a probability of 80 percent for 100 persons in the validation dataset, it would be expected that the actual number of deaths among those persons would be 80. However, there were only 54 instances of death by suicide in the validation dataset, and as a result, the actual probability was zero for many predicted probabilities. With an increased number of deaths by suicide in the modeling dataset, it is anticipated that the calibration variability would decrease. Even so, the calibration curve for the XGB model was less variable than the calibration curve for the logistic regression model, particularly for predicted probabilities higher than 0.5.

The models included class weights in order to assign equal importance to both outcome classes, and so the predicted probabilities were calibrated as though the modeling dataset contained balanced outcome classes. To evaluate the models calibrated to the risk of death by suicide in the modeling dataset, Platt calibration was used [22]. Platt calibration uses logistic regression to transform predicted probabilities into calibrated probabilities. Isotonic calibration was also tried but it achieved perfect calibration with the training dataset and poor calibration with the validation dataset. Figures 3 and 4 show the calibration curves for the logistic regression and XGB models, evaluated on both the training and validation datasets, and calibrated using Platt calibration. As before, the calibration curves were variable, and the calibration curve for the XGB model was less variable than the calibration curve for the logistic regression model, particularly for high predicted probabilities. For the Platt calibrated logistic regression model, predicted

probabilities below 0.2 appeared to be well calibrated but predicted probabilities above 0.2 appeared to be poorly calibrated. Similarly, the Platt calibrated XGB model appeared to be well calibrated except for predicted probabilities above 0.2, where the XGB model under-estimated the risk of death by suicide.

Platt calibration is commonly used to calibrate machine learning models because machine learning models often produce logistic s-shaped calibration curves. The models with outcome class weights did not produce logistic s-shaped calibration curves, and unfortunately, Platt calibration resulted in predicted probabilities of between 0.25 and 0.30 for all actual probabilities over 0.25. As an alternative to Platt calibration, a second XGB prediction model was developed to predict the actual probability of the outcome using the predicted probability of the outcome. The resulting calibration curves for the training and validation datasets (Figure 5) were better calibrated than the Platt calibration curves, although the validation dataset calibration curve was still variable.

4.5.3 Net Reclassification Improvement

The net reclassification improvement (NRI) for the optimal XGB model compared to the optimal logistic regression model using the models and the training and validation datasets from the calibration section above was 0.5183 (NRI_{event} = 0.6215, $NRI_{no\ event}$ = -0.1032) and 0.4644 (NRI_{event} = 0.5741, $NRI_{no\ event}$ = -0.1096) respectively.

4.5.4 Predictor Importance

The XGBoost library produces a measure of the importance of each predictor [23], and the 5 predictors with the highest importance from the optimal XGB model configuration (2 quarters of data, 70 classification trees, maximum tree depth of 2) were: the total number of emergency department visits with a parasuicide diagnosis that were classified as triage category 1 (from the most recent quarter); age (from the first quarter); the total number of inpatient days that were classified as maternity (from the most recent quarter); the suicide rate in the Local Geographic Area (community) of residence (from the most recent quarter); and the total cost of physician services (from the most recent quarter).

4.5.5 Tuning PPV using Class Weights

Clinicians are often interested in PPV because of its useful interpretation in clinical practice: the probability that a person will die by suicide given that they are identified as being at risk by the prediction model. The PPV of the optimal logistic regression model configuration was 0.0359 and the PPV of the optimal XGB model configuration was 0.0479. A higher PPV can be achieved by reducing the magnitude of the positive class weight configuration, with a decrease in sensitivity being the primary trade-off. The optimal XGB model configuration (2 quarters of data, 70 classification trees, maximum tree depth of 2) was evaluated with the full range of positive class weights from 1 to 125, and achieved a maximum 10-fold cross-validation PPV of 0.2016 using a class weight of 10, with sensitivity of 0.3686, specificity of 0.9884, and NPV of 0.9949. Higher 10-fold cross-validation PPV estimates were achieved with positive class weights below 10 but the estimates were highly variable across cross-validation folds.

4.6 Discussion

4.6.1 Implications for Model Development

The objective of this study is to compare the performance of logistic regression and XGB models that quantify the risk of death by suicide within 90 days of an ED visit for parasuicide using predictors available in administrative health care system data. It is unlikely that a single prediction model could be developed and implemented everywhere, and so researchers will likely be required to develop prediction models based on the administrative health care system data available to them.

The optimal XGB model configuration (AUC: 0.8786) displayed better discrimination than the optimal logistic regression model configurations with L2 regularization (AUC: 0.8632) and without L2 regularization (AUC: 0.8113). The optimal XGB model configuration also had better overall calibration (particularly following XGB calibration) than the optimal logistic regression model configuration with L2 regularization, particularly for persons at higher risk of death by suicide. The XGB calibration approach seems promising for calibrating machine learning models that do not produce logistic s-shaped calibration curves.

Both the optimal XGB and logistic regression model configurations achieved high 10-fold cross-validation AUC estimates which distinguishes these models from prior efforts to predict death by suicide. This could be because of the combination of predictors from a number of administrative data systems. For example, the 5 most important predictors in the optimal XGB model configuration each came from a different administrative data system: the total number of

emergency department visits with a parasuicide diagnosis that were classified as triage category 1, age, the total number of inpatient days that were classified as maternity, the suicide rate in the community of residence, and the total cost of physician services. It seems reasonable that each administrative data system would contribute to a fuller representation of each person, and this would provide prediction models with more information to make better predictions. The combination of personal and ecologic predictors could also be important for the high prediction performance. For example, the fourth most important predictor in the optimal XGB model configuration was the suicide rate in the community of residence.

An interesting finding from this study is that logistic regression without L2 regularization, which is the default in most statistical software, overfit to the training data and overfit more severely as the number of quarters increased. With the default scikit-learn L2 regularization enabled, the optimal logistic regression model configuration achieved a 10-fold cross-validation AUC estimate only slightly lower than the optimal XGB model configurations. This suggests that researchers that prefer logistic regression should consider regularization. Most statistical software includes procedures for regularization, although it might be referred to as ‘penalization’ or ‘shrinkage’.

Another interesting finding from this study that echoes previous work is that only the most recent 2 to 4 quarters (each with all 101 predictors) were needed for optimal performance. Performance increased as temporal data increased until a maximum was reached, after which additional temporal data resulted in decreasing or stationary performance. This suggests that

the risk state over the past year is most important for quantifying suicide risk in the current context.

Suicide risk and administrative data differs across jurisdictions, and researchers may need to develop their own prediction models rather than applying prediction models developed in other jurisdictions. Developing optimized models for implementation can be very costly and our studies were designed to provide readers with some direction by identifying the most promising classes of prediction models for quantifying suicide risk and determining the temporal period required for optimal performance. Future research should focus on obtaining as many instances of death by suicide as possible, and these instances may need to come from combining data across jurisdictions in order to obtain as many instances of death by suicide as possible. Future research should also focus on variable reduction to determine the minimal set required for optimal or near-optimal performance. For example, many predictors in the modeling dataset were never used for segmentation by the optimal XGB model configuration and would not be needed in an optimized production model. Predictor engineering is also likely to be important, particularly more refined diagnosis and intervention categories, and perhaps composite predictors.

In this study, good discrimination and calibration were achieved, and the performance seemed to be due more to the data than to the model classes. Although the calibrated XGB model demonstrated better discrimination than the calibrated logistic regression model, it could be argued that the improvement was incremental. The calibrated XGB model demonstrated a

material improvement in calibration compared with the calibrated logistic regression model, but still suffered from poor calibration for persons at highest risk. While the calibrated logistic regression model demonstrated high variability in the predicted probabilities for higher risk persons, the calibrated XGB model assigned a very narrow range of predicted probabilities for higher risk persons. The poor calibration for higher risk persons would likely be resolved with larger modeling datasets, particularly with more instances of death by suicide.

The goal of prediction modeling is to furnish health care service providers and health care policy providers with additional information to improve decisions. Prediction models that use administrative data would have access to information a clinician likely would not. For example, a clinician may not be able to access all health service records for a person presenting, and the person presenting may not be able to articulate the full details of their health services history. Further, one of the most important predictors in the optimal XGB model configuration was the suicide rate in the community of residence, and a clinician or person presenting may not be aware of the suicide rate in the community of residence. Also, even if a clinician had access to the same information as a prediction model, it would be unreasonable to expect the clinician to integrate the information into a superior risk estimate, and it has been shown that prediction models outperform clinicians.

In a sense, the utility of predicted probabilities would be to contribute to an informal Bayesian reasoning by clinicians. For example, when a person presents at an emergency department with parasuicide, a clinician would immediately be aware that this is a high-risk situation even before

meeting the person, which represents a pretest or prior probability of suicide risk. Then, the prediction model would provide a risk estimate, which may indicate a higher or lower risk. The clinician would update their pretest probability estimate, and meet the person presenting with a more refined prior probability. In meeting with the person presenting, the clinician would again update their probability estimate based on their clinical assessment, and make a better informed clinical judgment.

This study demonstrates that there is promise for realizing the above scenario to quantify the risk of death by suicide within 90 days of an ED visit for parasuicide, but to be clear, this study represents a step towards clinical innovation and not a recommendation for altered assessment. The calibrated XGB model configuration using a modeling dataset assembled from a number of administrative data systems demonstrated promising discrimination and calibration in a realistic health care setting. But whether furnishing clinicians with predicted probabilities actually leads to better clinical judgment requires further research. Poor predicted probabilities or good predicted probabilities that are integrated poorly have the potential to do harm. Once a prediction model is optimized for a particular clinical setting, clinical studies are necessary to determine how best to use the risk estimates in combination with clinical judgment. Then, once a model is implemented in clinical practice, clinical studies are necessary to determine if furnishing clinicians with predicted probabilities actually leads to better clinical judgment.

4.7 Limitations

There were three primary limitations in this study: the small number of instances of death by suicide in the modeling dataset, calibration assessment, and the inherent limitations of administrative data. The first two limitations are in a sense related because the Platt and XGB calibration curves seemed to be well calibrated overall but were variable, mainly because there were only 268 instances of death by suicide in the modeling dataset. With a larger number of instances of death by suicide it is anticipated that prediction models would result in calibration curves that would be smoother and better calibrated, particularly for persons with high actual risks of suicide.

The predictors available in the administrative data were not collected for the purposes of quantifying suicide risk and many important predictors were not available. Predictors that were not available in the administrative data but would be important would be direct measures of severity of mental illness, severity of substance misuse, suicidal ideation and intensity of suicidal ideation, social conditions and social interactions, and negative life events (death of a loved one, loss of employment, etc.). For example, while the number of health care services with a mental health diagnosis obtained from administrative health care system data can be a proxy for the severity of mental illness, a more direct measure of severity of mental illness would likely provide greater prediction utility. This is likely the most difficult limitation of administrative data to overcome, but this limitation could diminish if electronic health care system data becomes more complete and the ability to link with other data systems improves. Another limitation of developing prediction models with administrative data is that clinicians

would not be able to compute risk themselves and would have to rely on the development of an electronic application that would assemble predictors from multiple administrative data systems, quantify the risk of death by suicide using a prediction model, and provide a real-time, user-friendly interface to communicate the risk. Building such an electronic application and incorporating it into existing electronic medical record interfaces is not an impossible task, but the performance of the prediction model would have to warrant such an investment. This study is one of the first to show strong enough performance to warrant discussion about the feasibility of such an investment. We invite clinicians to consider and comment on the prediction performance required to justify such an investment, including preferred performance characteristics, such as the trade-off between PPV and sensitivity.

4.8 References

1. Alberta Vital Statistics. Cause of Death database; ICD-10: X60 through X84. 2019.

2. Mulder R, Newton-Howes G, Coid JW. The futility of risk prediction in psychiatry. Br J Psychiatry 2016; 209:271-272.

3. Large M, Kaneson M, Myles N, Myles H, Gunaratne P, Ryan C. Meta-analysis of longitudinal cohort studies of suicide risk assessment among psychiatric patients: heterogeneity in results and lack of improvement over time. PLoS ONE. 2016; 11(6):e0156322.

4. Huang X, Ribiero JD, Musacchio KM, Franklin JC. Demographics as predictors of suicidal thoughts and behaviors: A meta-analysis. PLoS ONE. 2017; 12(7):e0180793.

5. Chan MK, Bhatti H, Meader N, Stockton S, Evans J, O'Connor RC, et al. Predicting suicide following self-harm: systematic review of risk factors and risk scales. Br J Psychiatry. 2016; 209(4):277-83.

6. Carter G, Milner A, McGill K, Pirkis J, Kapur N, Spittal MJ. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. Br J Psychiatry. 2017.

7. Ribeiro JD, Franklin JC, Fox KR, Bentley KH, Kleiman EM, Chang BP, Nock MK. Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. Psychological Medicine. 2016; 46:225–236.

8. Saunders K, Brand F, Lascelles K, Hawton K. The sad truth about the SADPERSONS Scale: an evaluation of its clinical utility in self-harm patients. Emerg Med J. 2014; 31(10):796-8.

9. Katz C, Randall JR, Sareen J, et al. Predicting suicide with the SAD PERSONS scale. Depress Anxiety. 2017; 34(9):809-16.

10. Tran T, Luo W, Phung D, Harvey R, Berk M, Kennedy RL, Venkatesh S. Risk Stratification Using Data From Electronic Medical Records Better Predicts Suicide Risks Than Clinician Assessments. BMC Psychiatry. 2014; 14: 76.

11. Pisani AR, Murrie DC, Silverman MM. Risk Stratification Using Data from Electronic Medical Records Better Predicts Suicide Risks Than Clinician Assessments. Academic Psychiatry. 2016; 40: 623-629.

12. Sanderson M, Bulloch A, Wang J, Williamson T, Patten S. Predicting Death by Suicide Using Administrative Health Care System Data: Can Feedforward Neural Network Models Improve Upon Logistic Regression Models? Journal of Affective Disorders. 2019; 257:741-747.

13. Sanderson M, Bulloch A, Wang J, Williamson T, Patten S. Predicting Death by Suicide Using Administrative Health Care System Data: Can Recurrent Neural Network, One-Dimensional Convolutional Neural Network, and Gradient Boosted Trees Models Improve Prediction Performance? Journal of Affective Disorders. 2020; 264:107-114.

14. Olfson M, Marcus SC, Bridge JA. Focusing suicide prevention on periods of high risk. JAMA. 2014; 311(11):1107-8.

15. Alberta Health: Overview of Administrative Health Datasets. 2017.
https://open.alberta.ca/dataset/overview-of-administrative-health-datasets
Accessed on January 14, 2020.

16. Karmakar C, Luo W, Tran T, Berk M, Venkatesh S. Predicting risk of suicide attempt using history of physical illnesses from electronic medical records. JMIR Mental Health. 2016; 3:3.

17. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Olivier Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Journal of Machine Learning Research. 2011; 12(Oct):2825−2830.

18. Chen T, Guestrin C. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

19. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning, with Applications in R. 6th Printing, 2015. Springer, New York.

20. Hanley JA, McNeil BJ. The meaning and use of the area under a receiving operating characteristic (ROC) curve. Radiology. 1982; 143:29-36.

21. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html Accessed on January 14, 2020.

22. https://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html Accessed on January 14, 2020.

23. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html#sklearn.ensemble.GradientBoostingClassifier.feature_importances_ Accessed on January 14, 2020.

4.9 Appendix A: Table 1

Table 1: 10-Fold Cross-Validation Performance Metrics, Mean

Performance Metric	**Log. Regression Mean (1 Quarter)**	**Log. Regression L2 Mean (2 Quarters)**	**XGB Mean (2 Quarters)**
Area Under the Curve	0.8113	0.8632	0.8786
Accuracy	0.8531	0.8411	0.8895
Balanced Accuracy	0.7628	0.7925	0.7894
Sensitivity	0.6710	0.7429	0.6876
Specificity	0.8547	0.8420	0.8912
Positive Prediction Value	0.0354	0.0359	0.0479
Negative Prediction Value	0.9969	0.9974	0.9971

Figure 1: Calibration Curve, Logistic Regression, Predicted.

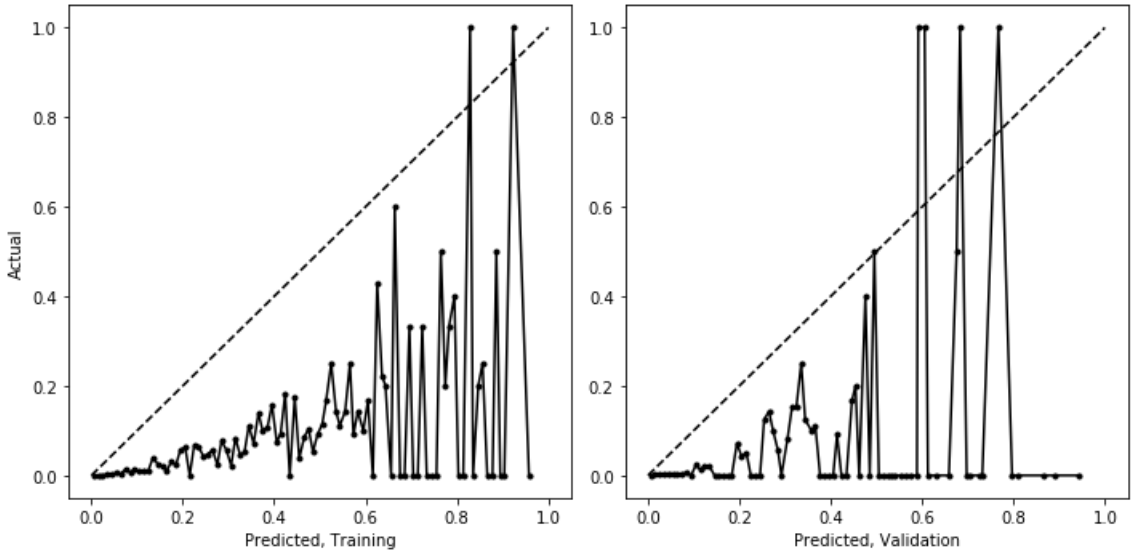


Figure 2: Calibration Curve, Gradient Boosted Trees, Predicted.

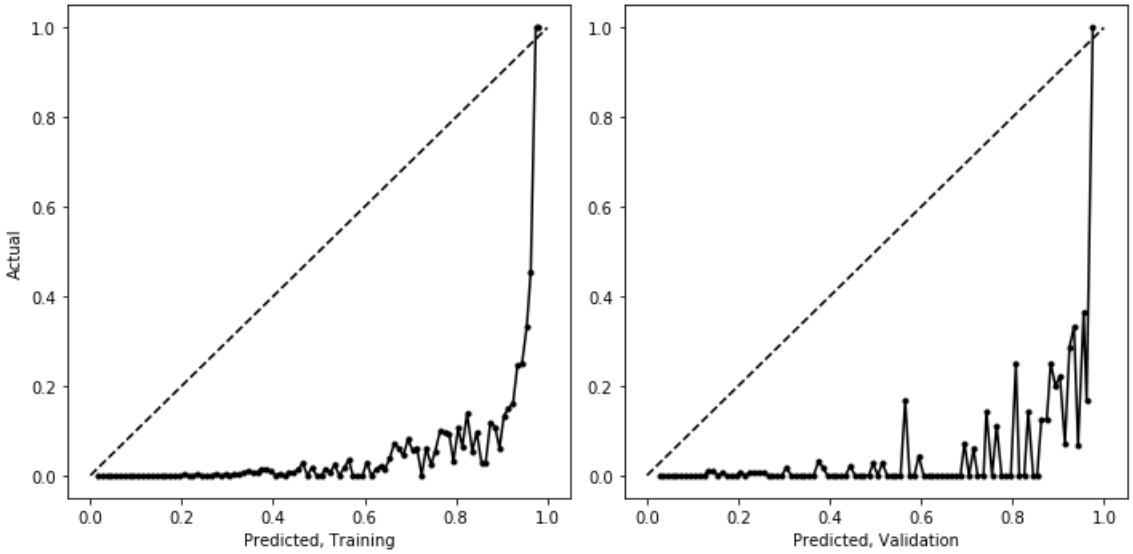


Figure 3: Calibration Curve, Logistic Regression, Platt Calibration.

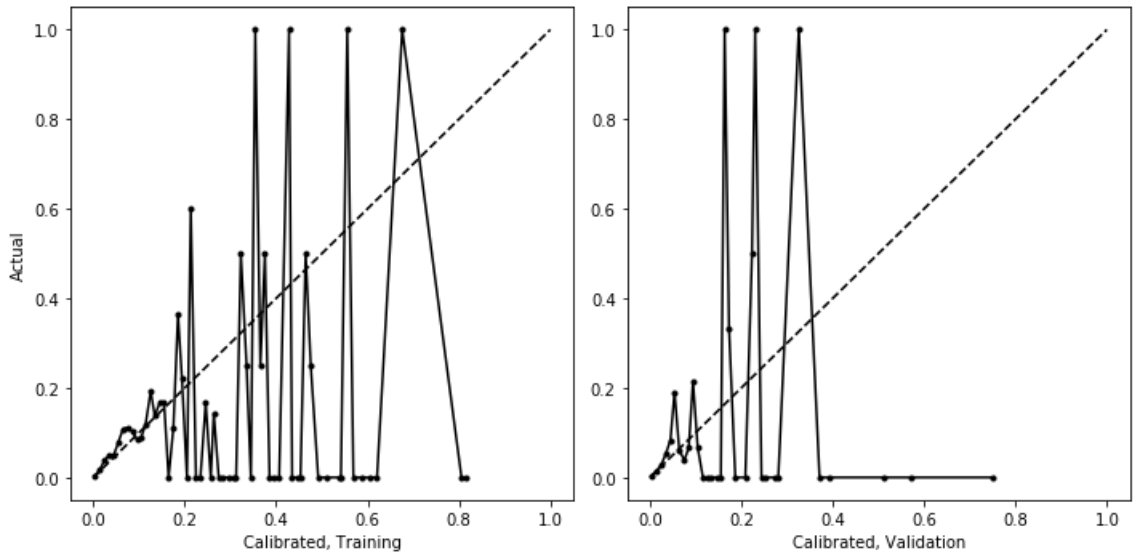


Figure 4: Calibration Curve, Gradient Boosted Trees, Platt Calibration.

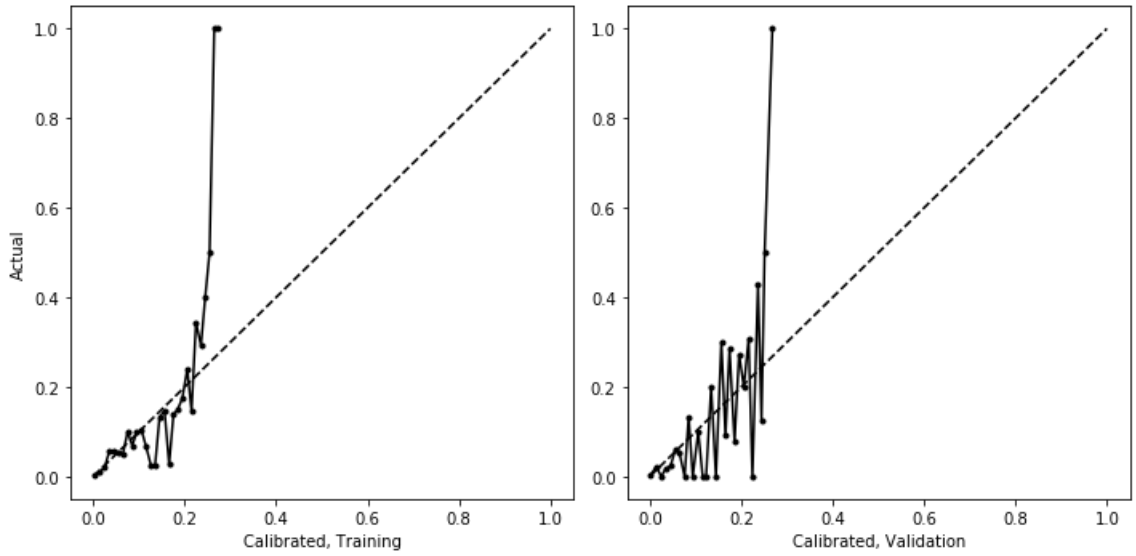
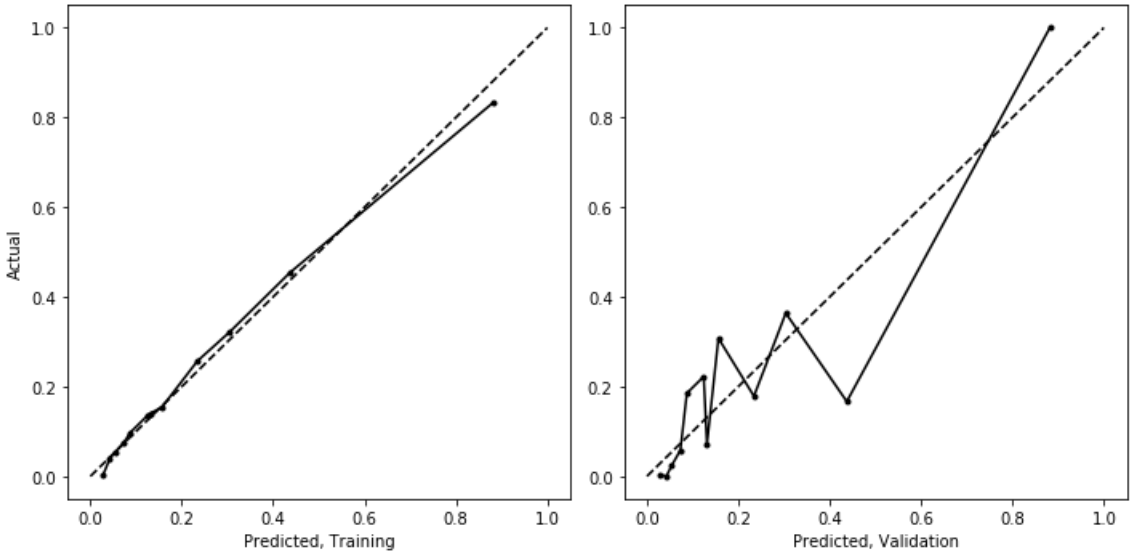


Figure 5: Calibration Curve, Gradient Boosted Trees, XGB Calibration.



CHAPTER 5: CONCLUSIONS

5.1 Contributions

The modeling datasets prepared for this thesis contained a larger volume of administrative health care system data than previous studies identified in the literature review in terms of the number of predictors, the time period covered by the predictors, and the number of administrative data sources. This thesis also evaluated a number of advanced prediction model classes that were not applied in the studies identified in the literature review. This thesis also included a robust model performance validation method (10-fold cross-validation), while most previously published studies did not validate model performance. This thesis showed that suicide prediction models developed with administrative health care system data assembled from a number of different data systems can achieve performance that could lead to clinical applications.

The first manuscript showed that administrative health care system data assembled from a number of data sources can achieve promising performance compared with the studies in the literature review, and that feedforward neural networks can outperform logistic regression. The second manuscript showed that cutting-edge machine learning model classes (recurrent neural networks, one-dimensional convolutional neural networks, and gradient boosted trees) can outperform the model classes in the first manuscript, and that 8 quarters of data or less is required for optimal performance. The third manuscript showed that prediction models developed with administrative health care system data can achieve promising performance in a clinical setting.

In the literature review, it had already been shown that suicidality prediction models developed with administrative health care system data generally outperformed risk scales and clinical judgment. However, these prediction models were developed with a smaller volume of administrative data and a smaller set of prediction model classes than those in this thesis. The studies in the literature review reported AUCs that were typically between 0.65 and 0.75, and always lower than 0.80. The AUC estimates in the first two manuscripts in this thesis, which used a case-control study design, were between 0.82 and 0.85. The AUC estimates of the two models in the third manuscript, which used a retrospective open-cohort study design, were 0.86 and 0.88. Further, the estimates in this thesis were validated using 10-fold cross validation, while most estimates in the literature review were not validated and are likely overestimates of their performance with unseen data.

In the second manuscript, Platt calibration resulted in essentially the same predicted risk for any actual risk over 0.70, while isotonic calibration severely overfit to the training data. In the third manuscript, Platt calibration resulted in essentially the same predicted risk for any actual risk over 0.20, while isotonic calibration again severely overfit to the training data. The third manuscript explored an alternative to Platt calibration and isotonic calibration to calibrate the predicted probabilities from the logistic regression and XGB models with class weights to the actual probabilities in the modeling dataset. While conceptually similar to Platt calibration, this approach used a XGB model to calibrate the predicted probabilities to actual probabilities. Although this approach achieved improvements in calibration in the third manuscript, the calibration curves were variable because of the small number of instances of the outcome.

However, this approach achieved clearly superior calibration with the modeling dataset from the second manuscript compared to Platt and isotonic calibration, and it is expected that with a larger number of instances of the outcome in the third manuscript, this approach would achieve similarly superior calibration.

Analytic prediction models have been adopted in clinical practice, and the most common statistical risk prediction model is the Framingham Risk Score (FRS) [15, 16, 17, 18]. The FRS has been widely adopted in clinical practice, and it estimates the 10-year risk of cardiovascular disease (CVD). A person's age, high-density lipoprotein (HDL) cholesterol, total cholesterol, systolic blood pressure, smoking status, and diabetes status are used to assign CVD risk as 'high' (20 percent or higher), 'intermediate' (10 to 19 percent), and 'low' (less than 10 percent). The FRS comes as a single sheet of paper, where a clinician can compute the FRS. The FRS also includes treatment targets based on risk categories. While it is unlikely that a prediction model that quantified the risk of death by suicide would be computable on a single sheet of paper like the FRS, the crucial next step for further research would be to determine what performance and usability characteristics are necessary for adoption in clinical practice.

Two notable psychiatric risk prediction models are the READMIT clinical risk index and the PredictD algorithm. The READMIT clinical risk index [19] was developed in Ontario to predict readmission within 30 days after discharge from acute psychiatric units with administrative health care system data. Although the READMIT clinical risk index was not developed to predict suicide risk, it used administrative health care system data to predict a psychiatric outcome in

persons receiving treatment for a psychiatric event, similar to the third manuscript. The READMIT index achieved moderate discrimination in the training (AUC: 0.631) and validation (AUC: 0.630) datasets. The PredictD algorithm [20] was developed to predict the onset of major depression with predictors obtained from a risk scale. The algorithm achieved an AUC of 0.79 in the training datasets and an AUC of 0.71 in an external validation dataset.

Although the manuscripts that comprise this thesis are not directly comparable to the studies in the literature review, the FRS, the READMIT clinical risk index, or the PredictD algorithm, the performance of the models in this thesis exceed existing suicide risk prediction models and the FRS, READMIT, and PredictD prediction models. Further, the manuscripts in this thesis were not designed to develop a finalized prediction model optimized for implementation in clinical practice, but rather to describe best practices for developing prediction models.

5.2 Further Research

This thesis took the early steps towards the ultimate goal of developing clinical applications that predict death by suicide using administrative health care system data, and showed that further research is warranted. The performance of the prediction models in the third manuscript demonstrated materially better performance than the studies in the literature review, and the performance was validated. While further research is needed, the importance of administrative health care system data and machine learning model classes to quantify suicide risk appears to be clear. The below sections describe next steps for further research.

5.2.1 Clinical Suitability

There is no consensus on what performance characteristics (sensitivity, specificity, PPV, NPV, calibration, etc.) would make prediction models attractive to clinicians. As described above, statistical prediction models have already been developed that outperform clinicians, but these models have not been widely adopted in clinical practice.

The prediction performance reported in this thesis exceeds the performance of those statistical models, but it is not known whether this performance is suitable for clinical practice. It may turn out that prediction models using administrative health care system data cannot achieve the necessary performance characteristics. For example, in the third manuscript, model configurations were adjusted to increase PPV at the expense of sensitivity, but it may be the case that a suitable trade-off between PPV and sensitivity cannot be achieved.

Determining clinical suitability may require studies of how clinicians make treatment decisions when faced with parasuicide, such as how clinicians make decisions given the probability of making different types of errors. Clinicians are required to make treatment decisions knowing that inpatient hospitalization could be invasive and potentially harmful for persons with a low risk of suicide, while discharge with routine follow-up could lead to death for persons with a high risk of suicide. One approach to achieving consensus on preferred performance characteristics could be a qualitative evaluation of clinician discussion and debate. While this discussion and debate has taken place to some degree in the literature, it has been focused on whether preferred performance characteristics could be achieved in principle, with little

discussion of what the preferred performance characteristics are. Another approach could be to formally conduct a cost-benefit analysis to quantitatively weigh the estimated costs and benefits of candidate performance characteristics. Another, more ambitious approach could be to present clinicians with simulated or retrospective clinical scenarios to understand how clinicians integrate prediction model output, how that might change the course of treatment, and quantitatively evaluate how that might impact suicide risk.

In addition, clinicians may be hesitant to adopt prediction models for reasons beyond performance considerations, including legal, ethical, and professional insurance considerations. For example, determining neglect or fault in a wrongful death claim could become much more complex. Currently, the standard against which clinical negligence is assessed is that of an acceptable standard of care, as judged by peers. If prediction models were used in clinical settings, it could be argued in a claim that a clinician gave too little or too much weight to the prediction model output when forming their clinical judgment. It could be discovered that a prediction model performs worse than expected in a particular group of persons, and argued that the model developers were negligent. For a prediction model that uses administrative data, it could be argued that a clinical coder negligently miscoded clinical information that would have otherwise allowed the prediction model and clinician to perform optimally. It could even be argued that the original clinical information coded by a clinical coder was negligent.

Of course, it must be acknowledged that most medical tests are imperfect. Their value lies in supplementing clinical judgment, despite being imperfect. Prediction models that quantify the

risk of death by suicide have not widely been used to supplement clinical judgment, likely because they have not achieved sufficient prediction performance to warrant implementation. The results reported in this thesis suggest that there is promise for prediction models developed with administrative health care systems data to achieve performance sufficient to supplement clinical judgment. If so, the standard of care (as mentioned above), which currently prioritizes clinical judgement, may need to change as the ability to quantify the risk of death by suicide improves.

Expending the large amount of resources required for the development of prediction models could be wasteful if clinicians are resistant (or even unable) to adopt prediction models in clinical practice. Determining whether prediction models that quantify the risk of death by suicide are likely to be suitable for clinical practice is the critical next step in determining whether further research could lead to clinical applications.

5.2.2 Prediction Model Optimization

Supposing that further research continues to confirm that prediction models are likely to be suitable for clinical practice, the next step for further research would be to optimize prediction model performance for a particular clinical setting. In addition to the methods described in this thesis for determining the optimal model configuration and temporal data requirements, a number of other methods bear consideration, such as additional methods to handle outcome class imbalance and methods to reduce the number of predictors in the modeling dataset.

When a modeling dataset contains an outcome class imbalance, models generally predict the majority class better than the minority class (although, this is dependent on a number of factors). This is a problem for modelers because the minority class is often the class of greatest interest. In this thesis, outcome class weights were used to handle the outcome class imbalance. In addition to outcome class weights, there are a number of other methods for imbalanced classification [14]. Several methods were tentatively evaluated and are described below.

One method is known as 'majority class undersampling'. With this approach, observations from the majority class are decreased so that the number of observations in the modeling dataset that belong to the minority and majority classes are more similar. The majority class can be decreased randomly, or can be decreased by removing selected observations. For example, two undersampling methods were tested as preliminary prediction model optimization methods with the modeling dataset from the third manuscript: undersampling with edited nearest neighbors (ENN) and undersampling with Tomek links. ENN removes an observation from the majority class if its nearest neighbors tend to belong to the minority class. Tomek links identify pairs of observations that belong to different outcome classes and are each other's nearest neighbors, and then removes the observation that belongs to the majority class. The concept behind these methods is that removing observations from the majority class will decrease the class imbalance and may improve prediction performance by removing majority class observations that are near decision boundaries. Both undersampling approaches resulted in

lower performance than using class weights with the modeling dataset from the third manuscript.

Another method used to handle outcome class imbalance is known as 'minority class oversampling'. With this approach, observations from the minority class are increased so that the number of observations in the modeling dataset that belong to the minority and majority classes are more similar. The minority class can be increased by duplicating observations, or can be increased by synthesizing observations. For example, two oversampling methods were tested as preliminary prediction model optimization methods with the modeling dataset from the third manuscript: borderline synthetic minority oversampling technique (borderline-SMOTE) and adaptive synthetic sampling (ADASYN). Borderline-SMOTE generates synthetic observations in the minority class by interpolating between minority class observations that are near majority class observations. ADASYN generates synthetic observations by interpolating between minority class observations (similar to SMOTE) but the number of observations generated is based on the proportion of nearest neighbors that belong to the minority class. The concept behind these methods is that synthesizing observations in the minority class will decrease the class imbalance and may improve prediction performance by synthesizing minority class observations that are near decision boundaries. Both oversampling approaches resulted in lower performance than using class weights with the modeling dataset from the third manuscript.

Another method to handle outcome class imbalance is known as ‘one-class learning’. This approach has been adapted from the field of anomaly detection. With this approach, observations from the minority class are treated as outliers to be detected relative to the majority class. One method of one-class learning called ‘One-Class Support Vector Machine’ uses a support vector machine to learn regions of the predictor space where the majority class has density. Observations that are outside of the regions of the predictor space where the majority class has density are considered outliers and are predicted to belong to the minority class. Another method of one-class learning is called ‘Isolation Forest’, and this approach uses an ensemble of decision trees to isolate observations that tend to be in terminal nodes nearer the root of the decision trees, which are then predicted to belong to the minority class. Another method of one-class learning is called ‘Local Outlier Factor’, and this approach uses k nearest neighbours. Observations that are far (Manhattan distance, Euclidean distance, or Minkowski distance) from their k nearest neighbours are considered outliers, and are predicted to belong to the minority class. Another method of one-class learning is called ‘Elliptic Envelope’, and it assumes that the predictors are Gaussian distributed. The majority class is used to estimate a multi-dimensional Gaussian ellipsoid in the predictor space, and observations outside the ellipsoid are considered outliers and are predicted to belong to the minority class. While one-class learning can be very effective when the minority class is distinct from the majority class, all of the above one-class learning approaches resulted in far lower performance than using class weights with the modeling dataset from the third manuscript.

Many of the above imbalanced learning methods rely on k nearest neighbours. When a modeling dataset contains a large number of predictors, methods based on k nearest neighbours often do not perform well because data points become sparse in high-dimensional space and a data point's nearest neighbours may not be similar to the data point. If the number of predictors in the modeling datasets in this thesis could be reduced, then methods based on k nearest neighbours could become more promising. In general, it is common in prediction modeling to seek to reduce the number of predictors in a modeling dataset without reducing prediction performance. Predictor reduction often requires both modeler and subject matter expert input, but this can be difficult with a large number of predictors.

Two automated predictor reduction methods were tested as preliminary prediction model optimization methods with the modeling dataset from the third manuscript: principal components analysis (PCA) and recursive predictor elimination (RPE). PCA reduces a larger set of predictors to a smaller set of predictors by combining correlated predictors into a single predictor. RPE reduces a larger set of predictors to a smaller set of predictors by recursively removing a selected number (often one at a time) of least important features until a desired number of predictors is achieved. PCA resulted in poorer performance than using all of the predictors in the modeling dataset from the third manuscript. RPE that removed 10 or less predictors at each recursive iteration was able to reduce the number of predictors without reducing prediction performance but at least 50 predictors were required in the reduced modeling dataset.

The performance of prediction models generally increases as the volume of data increases. Developing optimized prediction models will require a sufficient volume of data to produce optimal performance. With modeling datasets that contain an outcome class imbalance, the number of instances of the minority class is especially important. For example, in the calibration curves in the third manuscript, the predicted probability generally increased as the actual probability increased, but the agreement between the predicted and actual probabilities varied. It was anticipated that calibration variability would decrease with a larger number of deaths by suicide in the modeling dataset, and this was the case with the calibration curves in the second manuscript. The volume of data required to achieve optimal performance for a particular prediction model will differ depending on the clinical setting, and validation estimates will be important for determining the reliability of prediction model performance.

The methods described in this section were included as a preliminary exploration of prediction model optimization methods, and were meant to be illustrative and not exhaustive. Prediction model optimization is a dynamic field that relies on both modeler and subject matter expert insights, and often requires a large investment of analytic resources. The preliminary results of the methods described in this section reinforce that quantifying suicide risk is a complex problem and that convenient simplifications should not be expected. Determining the minimum set of predictors required to achieve optimal or near-optimal performance should be the first priority for prediction model optimization because fewer predictors would reduce model overfitting, would lessen the volume of data required to develop and implement prediction models, and would reduce the time required for additional performance optimization steps.

Predictor reduction would likely be costly in terms of computational expense, as well as modeler and subject matter expert time, and the trade-off between automated predictor reduction techniques and modeler and subject matter expert insights will be an important consideration. For example, while physical health has been shown to predict suicide in the absence of predictors related to mental health, predictors related to physical health in the modeling dataset in the third manuscript were not used by the optimal XGB model. Thus, predictors related to physical health appear to be candidates for predictor reduction but identifying other candidates may not be as easy.

It will also be important to investigate whether model performance varies across different groups of persons. If model performance does vary, it may be useful for this information to accompany the output so clinicians can temper their use of the prediction model output.

5.2.3 Prediction Model Implementation

Further research will be necessary to determine how best to use prediction model output in combination with clinical judgment. An important consideration would be whether training would be required to use the prediction model output in clinical practice. For example, in the third manuscript, it was suggested that the utility of prediction models would be to contribute to an informal Bayesian reasoning by clinicians, and so training might include heuristics describing how to apply prediction model output to Bayesian reasoning in circumstances likely to be encountered in clinical practice. Training could be optional, could be made a requirement of clinical licensure or specialty, or could be made a requirement of access to the prediction

model output. The appropriate training requirements could be determined based on further research.

Another consideration for prediction model implementation would be the format of the prediction model output, such as whether clinicians would prefer categorical output ('low risk' vs 'high risk', etc.) or a predicted probability or both. If a categorical output is preferred, the output categories could categorize risk and could also make treatment protocol recommendations ('standard evaluation' vs 'enhanced evaluation' vs 'hospitalize immediately', etc.). The type of output could even vary depending on the type of clinician. For example, if the clinician was a psychologist, the output categories and category risk thresholds may be different than if the clinician was a family physician.

Another consideration is the amount of control that a clinician would have over the performance characteristics of prediction models. For example, rather than being offered a single model, a clinician could be offered a set of models, each with different performance characteristics. Depending on the circumstances, a clinician may prefer the output of a particular prediction model that would better adjust their Bayesian prior risk estimate. A more ambitious concept would be a computer system that recomputes a prediction model based on the performance characteristics a clinician would like to maximize. It could even be possible that important predictors could be collected by clinicians and it would not be necessary to obtain them from administrative health care system data.

Another important consideration would be whether the prediction model outputs would strictly be used as supplementary information or whether certain outputs would require a mandatory treatment protocol. A similar consideration is whether prediction model outputs would routinely accompany an electronic medical record or would only be computed when requested by a clinician (similar in mechanism to a laboratory requisition). Providing model outputs in all clinical settings could lead to over-vigilance, while providing model outputs only when requisitioned could lead to under-vigilance. Future research would be needed to determine the optimal availability of prediction model outputs.

It is expected that it is currently possible to develop a computer system to assemble the electronic administrative health care system records for a person, input the data into a prediction model, and deliver the output in real time. It is also expected that developing such a computer system would be challenging due to the fractured nature of clinical electronic information systems (legacy systems, differing electronic medical record platforms, etc.). The computer system interface would be vital to successful clinical implementation, and the focus should be on ensuring that the interface enhances clinical practice and does not interfere with clinical practice. Qualitative research, such as usability testing in simulated clinical settings, could be required discover how to best present risk estimates in a computer interface.

Regardless of the prediction model output and computer interface chosen, a formal randomized controlled trial (RCT) would be necessary prior to widespread implementation. The RCT would be designed to evaluate whether the preferred performance characteristics were

achieved, as well as to provide insights into standard clinical considerations (such as safety) and to identify unanticipated difficulties.

5.2.4 Prediction Model Implementation Evaluation

If prediction models that quantify the risk of death by suicide using administrative health care system data become implemented in clinical practice, then research will be needed to evaluate whether the goals of implementing the prediction models were achieved. Research would be required to determine whether the expected performance of the prediction model as estimated by k-fold cross-validation during model development and as estimated in the RCT (described above) was actually achieved when implemented.

Research would also be required to discover whether clinicians found the prediction model output useful in clinical practice. For example, qualitative research could evaluate whether clinicians found that the output was useful or found that the inability of the prediction model to directly describe what an elevated risk was attributable to was a hindrance. It would also be valuable to discover what refinements clinicians would recommend for the prediction model and the computer interface.

Whether suicide prediction models developed with administrative health care system data will become part of routine clinical practice and will be emulated in other health care settings will be decided by research that will evaluate whether providing clinicians with prediction model output actually leads to better clinical judgment, and whether the improvement justifies the

cost and effort. If so, then a cycle of continuous improvement of the prediction models, computer interfaces, and possibly data sources, would begin. This cycle may eventually lead to the development of prediction models that include administrative health care system data, clinical assessment tool data, and clinical judgment as input. It may even lead to prediction models that also include physiological predictors (blood tests, galvanic skin response, genetic markers, etc.) as input.

5.2.5 Other Jurisdictions

If providing clinicians with prediction model outputs leads to better clinical judgment, other jurisdictions may be interested in developing prediction models. It is unlikely that a prediction model developed with administrative health care system data in one jurisdiction could be directly applied in another jurisdiction. This is because the administrative health care system data in the jurisdictions would likely be different. Research could focus on using data from another jurisdiction as input to a prediction model but it is likely that focusing on applying research to develop prediction models from scratch (this thesis, for example) in other jurisdictions would be more fruitful. Research could also focus on whether it would be feasible for jurisdictions to develop a common prediction model based on a minimum set of predictors that all jurisdictions could assemble, even if the performance is less than optimal.

The particular models developed for this thesis may not be directly applicable in another jurisdiction but the prediction modeling approach would be directly applicable. For example, the improvements in discrimination, calibration, and computational expense achieved by

gradient boosted trees models compared with logistic regression and neural networks is likely to be replicated in other jurisdictions, as well as the approaches to calibrate the predicted probabilities from prediction models. Similarly, the recency and breadth of administrative health care system data required for optimal performance is likely to be applicable in other jurisdictions.

5.2.6 Summary

Pessimism has long been expressed about the possibility of quantifying the risk of death by suicide for clinical applications, but this thesis suggests that the near future should be a time of optimism. This thesis combined administrative health care system data from five different administrative data systems with advances in prediction model classes and the increased availability of computer software and hardware to non-specialists, and achieved promising prediction performance. The key contribution of this thesis was to demonstrate that there is promise for quantifying the risk of death by suicide for use in clinical applications, and to provide foundations and directions for future research.

CHAPTER 6: REFERENCES AND BIBLIOGRAPHY

6.1 References

1. Leading Causes of Death in Canada: Alberta. Statistics Canada. 2009.

http://www.statcan.gc.ca/pub/84-215-x/2012001/tbl/T022-eng.pdf

2. Preventing Suicide: A Global Imperative. World Health Organization. 2014.

http://apps.who.int/iris/bitstream/10665/131056/1/9789241564779_eng.pdf?ua=1&ua=1

3. Alberta Vital Statistics. Cause of Death database; ICD-10: X60 through X84. 2019.

4. Alberta Health: Supplemental Enhanced Service Event (SESE) physician claims database. 2019.

5. Niederkrotenthaler T, Sonneck G, Dervic K, Nader IW, Voracek M, Kapusta ND, Etzersdorfer E,Mittendorfer-Rutz E, Dorner T. Predictors of Suicide and Suicide Attempt in Subway Stations: A Population-based Ecological Study. Journal of Urban Health. 2012; 89: 339-353.

6. Pisani AR, Murrie DC, Silverman MM. Risk Stratification Using Data from Electronic Medical Records Better Predicts Suicide Risks Than Clinician Assessments. Academic Psychiatry. 2016; 40: 623-629.

7. Wang Y, Bhaskaran J, Sareen J, Bolton S, Chateau D, Bolton JM. Clinician Prediction of Future Suicide Attempts. Canadian Journal of Psychiatry. 2016; 61: 428-432.

8. Tran T, Luo W, Phung D, Harvey R, Berk M, Kennedy RL, Venkatesh S. Risk Stratification Using Data From Electronic Medical Records Better Predicts Suicide Risks Than Clinician Assessments. BMC Psychiatry. 2014; 14: 76.

9. Yaseen ZS, Kopeykina I, Gutkovich Z, Bassirnia A, Cohen LJ, Galynker II. Predictive Validity of the Suicide Trigger Scale (STS-3) for Post-Discharge Suicide Attempt in High-Risk Psychiatric Inpatients. PLoS ONE. 2014; 9.

10. Karmakar C, Luo W, Tran T, Berk M, Venkatesh S. Predicting Risk of Suicide Attempt Using History of Physical Illnesses From Electronic Medical Records. JMIR Mental Health. 2016; 3: 3.

11. Bae SM, Lee SA, Lee S. Prediction by Data Mining, of Suicide Attempts in Korean Adolescents: a National Study. Neuropsychiatric Disease and Treatment. 2015; 11: 2367-2375.

12. Poulin C, Shiner B, Thompson P, Vepstas L, Young-Xu Y, Goertzel B, Watts B, Flashman L, McAllister T. Predicting the Risk of Suicide by Analyzing the Text of Clinical Notes. PLoS ONE. 2014; 9: 1.

13. Cook BL, Progovac AM, Chen P, Mullin B, Hou S, Baca-Garcia E. Novel Use of Natural Language Processing (NLP) to Predict Suicidal Ideation and Psychiatric Symptoms in a Text-Based Mental Health Intervention in Madrid. Computational and Mathematical Methods in Medicine. 2016.

14. https://imbalanced-learn.readthedocs.io/en/stable/api.html

15. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care. The Framingham Heart Study. Circ 2008; 117:743-53.

16. Genest J , McPherson R, Frohlich J, Anderson T, Campbell N, Carpentier A, Couture P, Dufour R, Fodor G, Francis GA, Grover S, Gupta M, Hegele RA, Lau DC, Leiter L, Lewis GF, Lonn E, Mancini GBJ, Ng D, Pearson GJ, Sniderman, Stone JA, Ur E. 2009 Canadian Cardiovascular

Society/Canadian guidelines for the diagnosis and treatment of dyslipidemia and prevention of cardiovascular disease in the adult. Can J Cardiol. 2009; 25(10):567-579.

17. Anderson T et al.(i). Anderson TJ , Grégoire J, Hegele RA, Couture P, Mancini GBJ, McPherson R, Francis GA, Poirier P, Lau DC, Grover S, Genest J, Carpentier AC, Dufour R, Gupta M, Ward R, Leiter LA, Lonn E, Ng DS, Pearson GJ, Yates GM, Stone JA, Ur E. 2012 Update of the Canadian Cardiovascular Society guidelines for the diagnosis and treatment of dyslipidemia for the prevention of cardiovascular disease in the adult. Can J Cardiol. 2013; 29(2):151-167.

18. https://www.ccs.ca/images/Guidelines/Tools_and_Calculators_En/FRS_eng_2017_fnl1.pdf

19. Vigod SN, Kurdyak PA, Seitz D, Herrmann N, Fung K, Lin E, Perlman C, Taylor VH, Rochon PA, Gruneir A. READMIT: a clinical risk index to predict 30-day readmission after discharge from acute psychiatric units. J Psychiatr Res. 2015;61:205-213. doi:10.1016/j.jpsychires.2014.12.003

20. King M, Walker C, Levy G, Bottomley C, Royston P, Weich S, Bellón-Saameño JA, Moreno B, Svab I, Rotar D, Rifel J, Maaroos H, Aluoja A, Kalda R, Neeleman J, Geerlings MI, Xavier M, Carraça I, Gonçalves-Pereira M, Vicente B, Saldivia S, Melipillan R, Torres-Gonzalez F, Nazareth I. Arch Gen Psychiatry. 2008 Dec;65(12):1368-76. doi: 10.1001/archpsyc.65.12.1368.

6.2 Bibliography

Alberta Health: Overview of Administrative Health Datasets. 2017.

http://www.health.alberta.ca/documents/Research-Health-Datasets.pdf

Alberta Vital Statistics. Cause of Death database; ICD-10: X60 through X84. 2019.

Allebeck P, Allgulander C, Fisher LD. Predictors of Completed Suicide in a Cohort of 50,465 Young Men: Role of Personality and Deviant Behaviour. British Medical Journal. 1988; 297: 176-178.

Anderson T et al.(i). Anderson TJ , Grégoire J, Hegele RA, Couture P, Mancini GBJ, McPherson R, Francis GA, Poirier P, Lau DC, Grover S, Genest J, Carpentier AC, Dufour R, Gupta M, Ward R, Leiter LA, Lonn E, Ng DS, Pearson GJ, Yates GM, Stone JA, Ur E. 2012 Update of the Canadian Cardiovascular Society guidelines for the diagnosis and treatment of dyslipidemia for the prevention of cardiovascular disease in the adult. Can J Cardiol. 2013; 29(2):151-167.

Bae SM, Lee SA, Lee S. Prediction by Data Mining, of Suicide Attempts in Korean Adolescents: a National Study. Neuropsychiatric Disease and Treatment. 2015; 11: 2367-2375.

Carter G, Milner A, McGill K, Pirkis J, Kapur N, Spittal MJ. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. Br J Psychiatry. 2017.

Chan MK, Bhatti H, Meader N, Stockton S, Evans J, O'Connor RC, et al. Predicting suicide following self-harm: systematic review of risk factors and risk scales. Br J Psychiatry. 2016; 209(4):277-83.

Chen T, Guestrin C. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

Cook BL, Progovac AM, Chen P, Mullin B, Hou S, Baca-Garcia E. Novel Use of Natural Language Processing (NLP) to Predict Suicidal Ideation and Psychiatric Symptoms in a Text-Based Mental Health Intervention in Madrid. Computational and Mathematical Methods in Medicine. 2016.

D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care. The Framingham Heart Study. Circ 2008; 117:743-53.

Fairweather-Schmidt AK, Anstey KJ, Salim A, Rodgers B. Baseline Factors Predictive of Serious Suicidality at Follow-Up: Findings Focussing on Age and Gender from a Community-Based Study. BMC Psychiatry. 2010; 10: 41.

Galynker I, Yaseen ZS, Briggs J, Hayashi F. Attitudes of Acceptability and Lack of Condemnation Toward Suicide May Be Predictive of Post-Discharge Suicide Attempts. BMC Psychiatry. 2015; 15: 87.

Genest J , McPherson R, Frohlich J, Anderson T, Campbell N, Carpentier A, Couture P, Dufour R, Fodor G, Francis GA, Grover S, Gupta M, Hegele RA, Lau DC, Leiter L, Lewis GF, Lonn E, Mancini GBJ, Ng D, Pearson GJ, Sniderman, Stone JA, Ur E. 2009 Canadian Cardiovascular Society/Canadian guidelines for the diagnosis and treatment of dyslipidemia and prevention of cardiovascular disease in the adult. Can J Cardiol. 2009; 25(10):567-579.

Glenn CR, Nock MK. Improving the Prediction of Suicidal Behavior in Youth. International Journal of Law and Psychiatry. 2013; 36: 390-398.

Gonda X, Fountoulakis KN, Kaprinis G, Rihmer Z. Prediction and Prevention of Suicide in Patients with Unipolar Depression and Anxiety. Annals of General Psychiatry. 2008; 7: S23.

Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press. 2016. 192-195.

Hanley JA, McNeil BJ. The meaning and use of the area under a receiving operating characteristic (ROC) curve. Radiology. 1982; 143:29-36.

Hornik K. (1991). Approximation Capabilities of Multilayer Feedforward Networks. Neural Networks. 1991; 4(2):251–257.

Horwitz AG, Czyz EK, King CA. Predicting Future Suicide Attempts Among Adolescent and Emerging Adult Psychiatric Emergency Patients. Journal of Clinical Child and Adolescent Psychology. 2014; 44: 751-761.

Huang X, Ribiero JD, Musacchio KM, Franklin JC. Demographics as predictors of suicidal thoughts and behaviors: A meta-analysis. PLoS ONE. 2017; 12(7):e0180793.

Ishtiak-Ahmed K, Perski A, Mittendorfer-Rutz E. Predictors of Suicidal Behaviour in 36,304 Individuals Sickness Absent Due To Stress-Related Mental Disorders - a Swedish Register Linkage Cohort Study. BMC Public Health. 2013; 13: 492.

James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning, with Applications in R. 6th Printing, 2015. Springer, New York.

Karmakar C, Luo W, Tran T, Berk M, Venkatesh S. Predicting Risk of Suicide Attempt Using History of Physical Illnesses From Electronic Medical Records. JMIR Mental Health. 2016; 3: 3.

Katz C, Randall JR, Sareen J, et al. Predicting suicide with the SAD PERSONS scale. Depress Anxiety. 2017; 34(9):809-16.

King M, Walker C, Levy G, Bottomley C, Royston P, Weich S, Bellón-Saameño JA, Moreno B, Svab I, Rotar D, Rifel J, Maaroos H, Aluoja A, Kalda R, Neeleman J, Geerlings MI, Xavier M, Carraça I, Gonçalves-Pereira M, Vicente B, Saldivia S, Melipillan R, Torres-Gonzalez F, Nazareth I. Arch Gen Psychiatry. 2008 Dec;65(12):1368-76. doi: 10.1001/archpsyc.65.12.1368.

Large M, Kaneson M, Myles N, Myles H, Gunaratne P, Ryan C. Meta-analysis of longitudinal cohort studies of suicide risk assessment among psychiatric patients: heterogeneity in results

and lack of improvement over time. PLoS ONE. 2016; 11(6):e0156322.

Leavey G, Rosato M, Galway K, Hughes L, Mallon S, Rondon J. Patterns and Predictors of Help-Seeking Contacts with Health Services and General Practitioner Detection of Suicidality Prior to Suicide: A Cohort Analysis of Suicides Occurring Over a Two-Year Period. BMC Psychiatry. 2016; 16: 120.

Mrnak-Meyer J, Tate SR, Tripp JC, Worley MJ, Jajodia A, McQuaid JR. Predictors of Suicide-Related Hospitalization among U.S. Veterans Receiving Treatment for Comorbid Depression and Substance Dependence. Who is the Riskiest of the Risky? Suicide & Life-Threatening Behavior. 2011; 41: 532-542.

Mulder R, Newton-Howes G, Coid JW. The futility of risk prediction in psychiatry. Br J Psychiatry 2016; 209:271-272.

Niederkrotenthaler T, Sonneck G, Dervic K, Nader IW, Voracek M, Kapusta ND, Etzersdorfer E,Mittendorfer-Rutz E, Dorner T. Predictors of Suicide and Suicide Attempt in Subway Stations: A Population-based Ecological Study. Journal of Urban Health. 2012; 89: 339-353.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Olivier Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Journal of Machine Learning Research. 2011; 12(Oct):2825−2830.

Pisani AR, Murrie DC, Silverman MM. Risk Stratification Using Data from Electronic Medical Records Better Predicts Suicide Risks Than Clinician Assessments. Academic Psychiatry. 2016; 40: 623-629.

Pisani AR, Murrie DC, Silverman MM. Reformulating Suicide Risk Formulation: From Prediction to Prevention. Acad Psychiatry. 2016 Aug; 40(4):623-9

Poulin C, Shiner B, Thompson P, Vepstas L, Young-Xu Y, Goertzel B, Watts B, Flashman L, McAllister T. Predicting the Risk of Suicide by Analyzing the Text of Clinical Notes. PLoS ONE. 2014; 9: 1.

Prediction and prevention of suicide. Canadian Medical Association Journal. 1969; 100: 867-868.

Prinstein MJ, Nock MK, Simon V, Aikins JW, Cheah CSL, Spirito A. Longitudinal Trajectories and Predictors of Adolescent Suicidal Ideation and Attempts Following Inpatient Hospitalization. Journal of Consulting and Clinical Psychology. 2008; 76: 92-103.

Ribeiro JD, Franklin JC, Fox KR, Bentley KH, Kleiman EM, Chang BP, Nock MK. Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. Psychological Medicine. 2016; 46:225–236.

Ribeiro JD, Franklin JC, Fox KR, Bentley KH, Kleiman EM, Chang BP, Nock MK. Suicide as a complex classification problem: machine learning and related techniques can advance suicide prediction – a reply to Roaldset. Psychological Medicine. 2016; 46:2009–2010.

Saunders K, Brand F, Lascelles K, Hawton K. The sad truth about the SADPERSONS Scale: an evaluation of its clinical utility in self-harm patients. Emerg Med J. 2014; 31(10):796-8.

Statistics Canada. Leading Causes of Death in Canada: Alberta. 2009. http://www.statcan.gc.ca/pub/84-215-x/2012001/tbl/T022-eng.pdf

Tianen BJ, Gustafson DH. A Comparison of Models for Predicting the Outcome of Suicide Attempts. Proceedings of the Annual Symposium on Computer Application in Medical Care. 1981; 398-405.

Ting SA, Sullivan AF, Miller I, Espinola JA, Allen MH, Camargo CA, Boudreaux ED. Multicenter Study of Predictors of Suicide Screening in Emergency Departments. Academic Emergency Medicine. 2012; 19: 239-243.

Tran T, Luo W, Phung D, Harvey R, Berk M, Kennedy RL, Venkatesh S. Risk Stratification Using Data From Electronic Medical Records Better Predicts Suicide Risks Than Clinician Assessments. BMC Psychiatry. 2014; 14: 76.

UNC Vision Lab. 2017.

http://image-net.org/challenges/LSVRC/2017/

Vigod SN, Kurdyak PA, Seitz D, Herrmann N, Fung K, Lin E, Perlman C, Taylor VH, Rochon PA, Gruneir A. READMIT: a clinical risk index to predict 30-day readmission after discharge from acute psychiatric units. J Psychiatr Res. 2015;61:205-213. doi:10.1016/j.jpsychires.2014.12.003

Wang Y, Bhaskaran J, Sareen J, Bolton S, Chateau D, Bolton JM. Clinician Prediction of Future Suicide Attempts. Canadian Journal of Psychiatry. 2016; 61: 428-432.

World Health Organization. Preventing Suicide: A Global Imperative. 2014.

http://apps.who.int/iris/bitstream/10665/131056/1/9789241564779_eng.pdf?ua=1&ua=1

Yaseen ZS, Kopeykina I, Gutkovich Z, Bassirnia A, Cohen LJ, Galynker II. Predictive Validity of the Suicide Trigger Scale (STS-3) for Post-Discharge Suicide Attempt in High-Risk Psychiatric Inpatients. PLoS ONE. 2014; 9.

Yen S, Shea MT, Walsh Z, Edelen MO, Hopwood CJ, Markowitz JC, Ansell EB, Morey LC, Grilo CM, Sanislow CA, Skodol AE, Gunderson JG, Zanarini MC, McGlashan TH. Self-Harm Subscale of the Schedule for Nonadaptive and Adaptive Personality (SNAP): Predicting Suicide Attempts Over 8 Years of Follow-Up. The Journal of Clinical Psychiatry. 2011; 72: 1522-1528.

https://www.ccs.ca/images/Guidelines/Tools_and_Calculators_En/FRS_eng_2017_fnl1.pdf

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

https://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html#sklearn.ensemble.GradientBoostingClassifier.feature_importances_

https://imbalanced-learn.readthedocs.io/en/stable/api.html

APPENDIX A: FIGURES

Figure 1. Suicide Rate in Alberta, 1983 – 2017

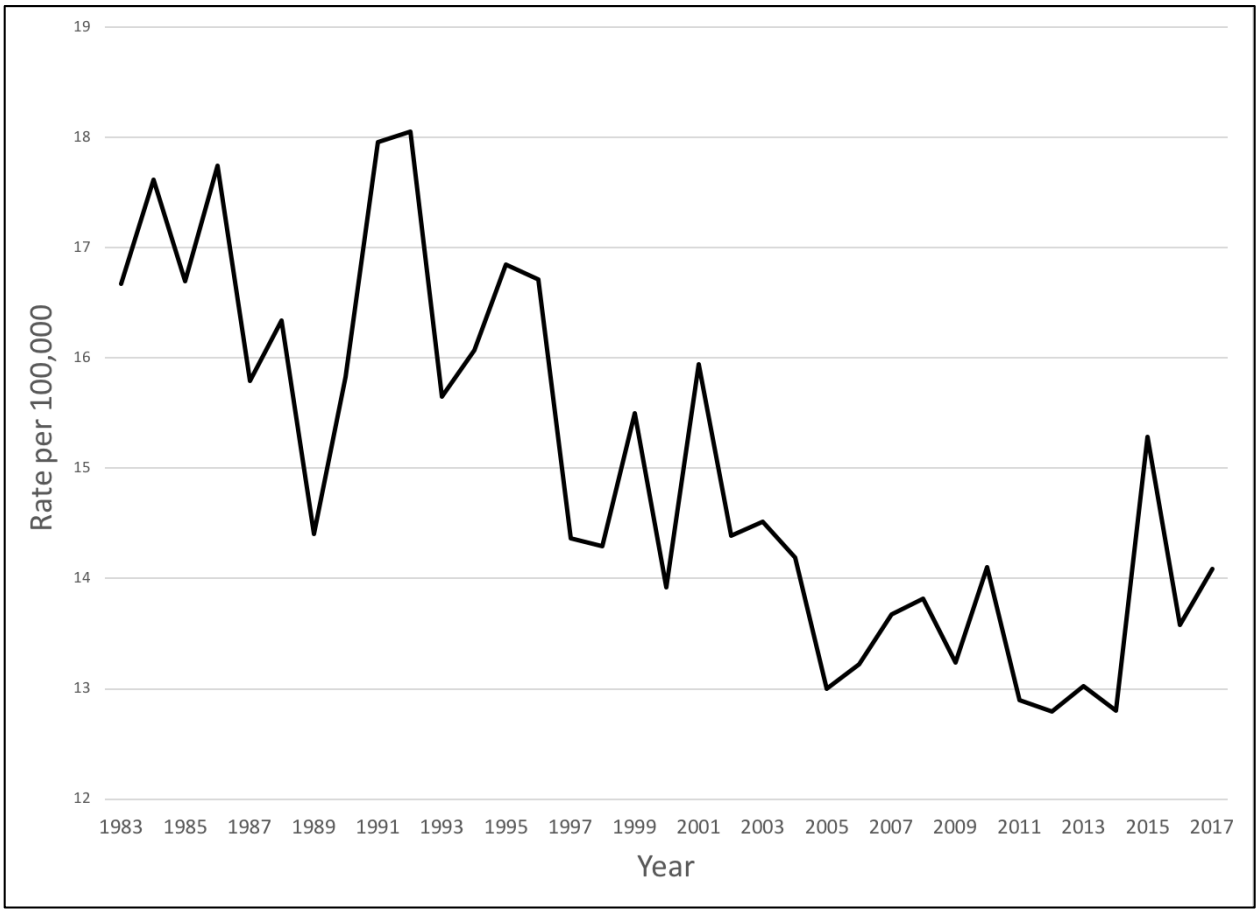


Figure 2. Suicide Rate in Alberta, Females and Males, 1983 – 2017

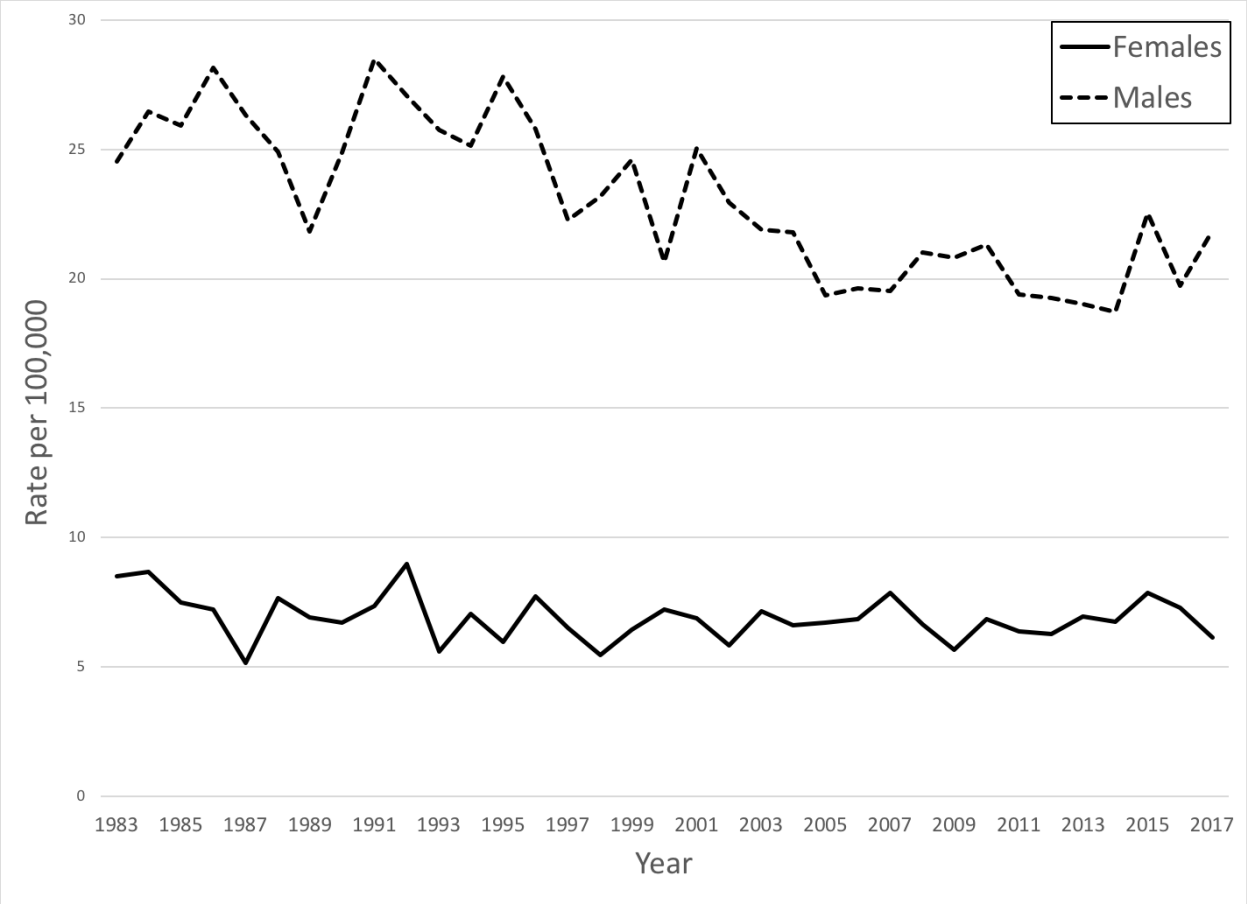


Figure 3. Suicide Rate in Alberta by Sex and Age, 2000 - 2017

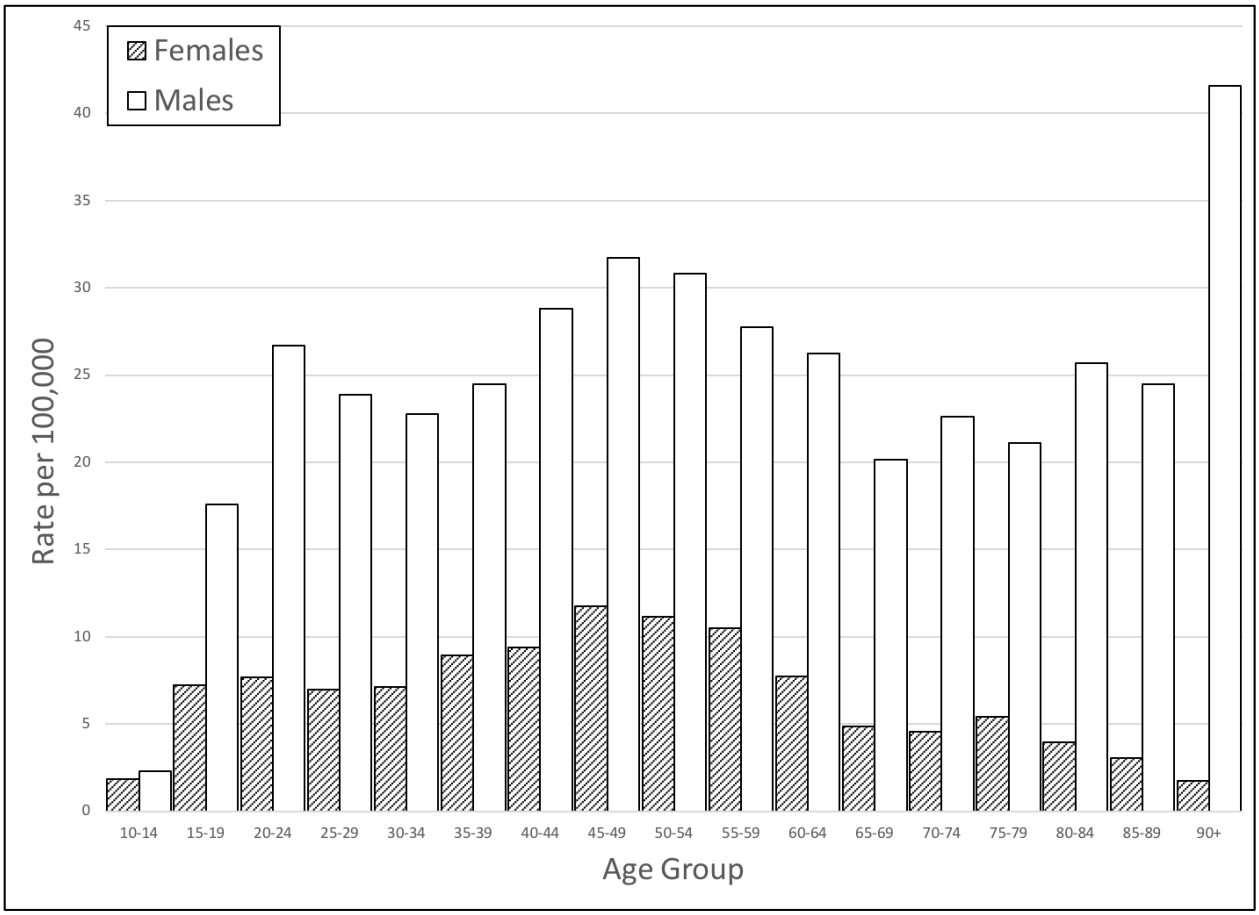


Figure 4. Suicide Rate in Alberta by Community, 2000 - 2017

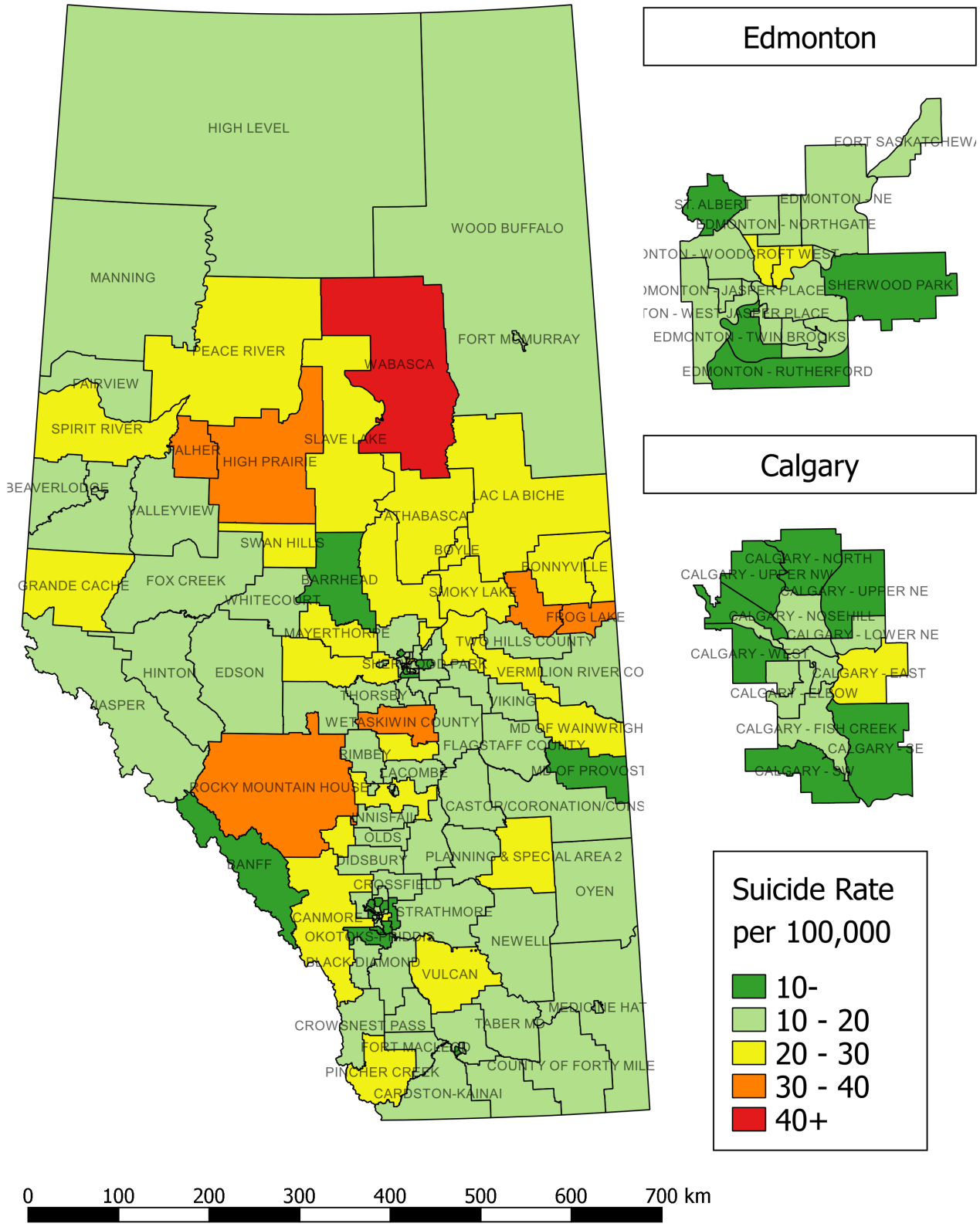


Figure 5. Suicide Rate in Alberta by Rurality, 2000 – 2017

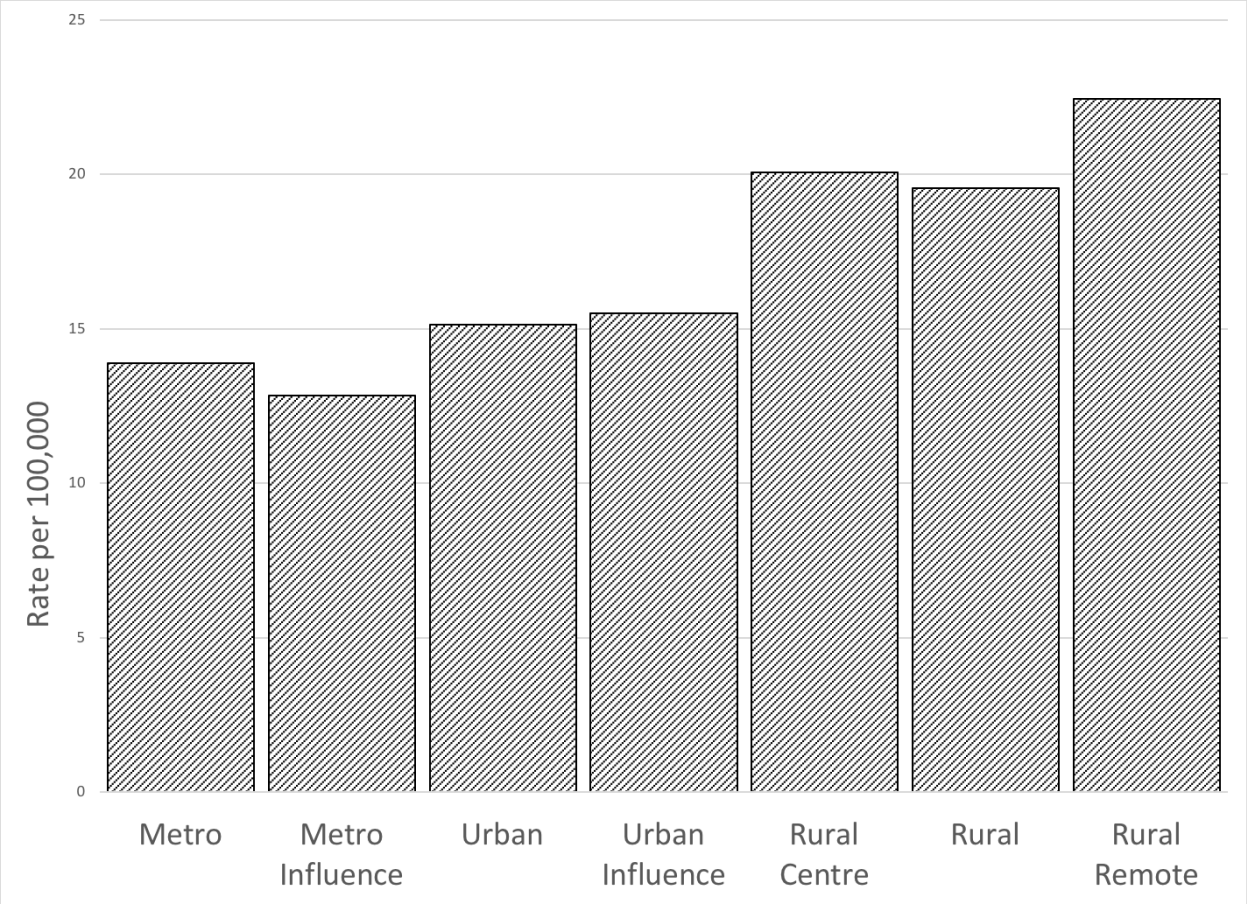


Figure 6. Scatterplot of the suicide rate and after-tax household income adjusted for household size (LIM-AT) rate in Alberta, 2016

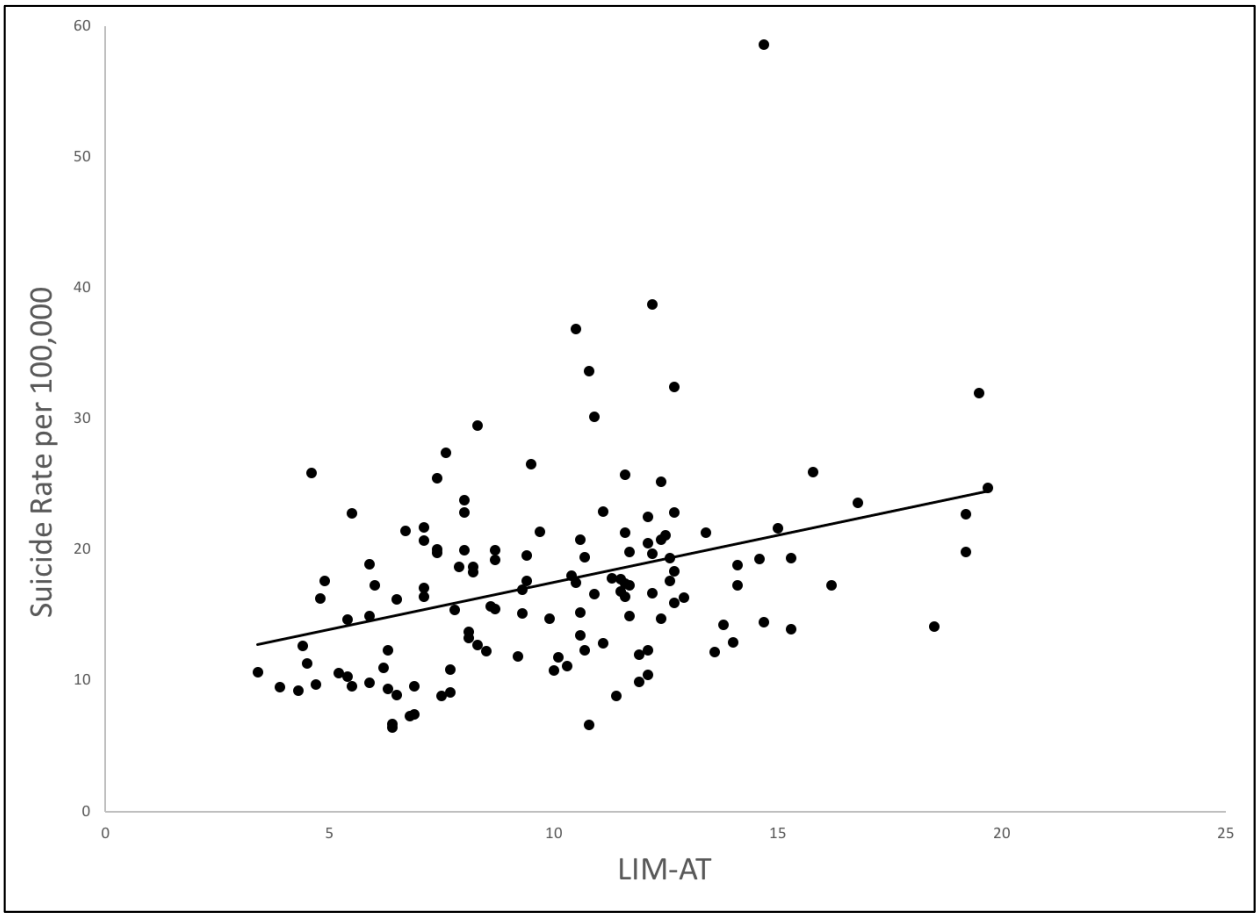
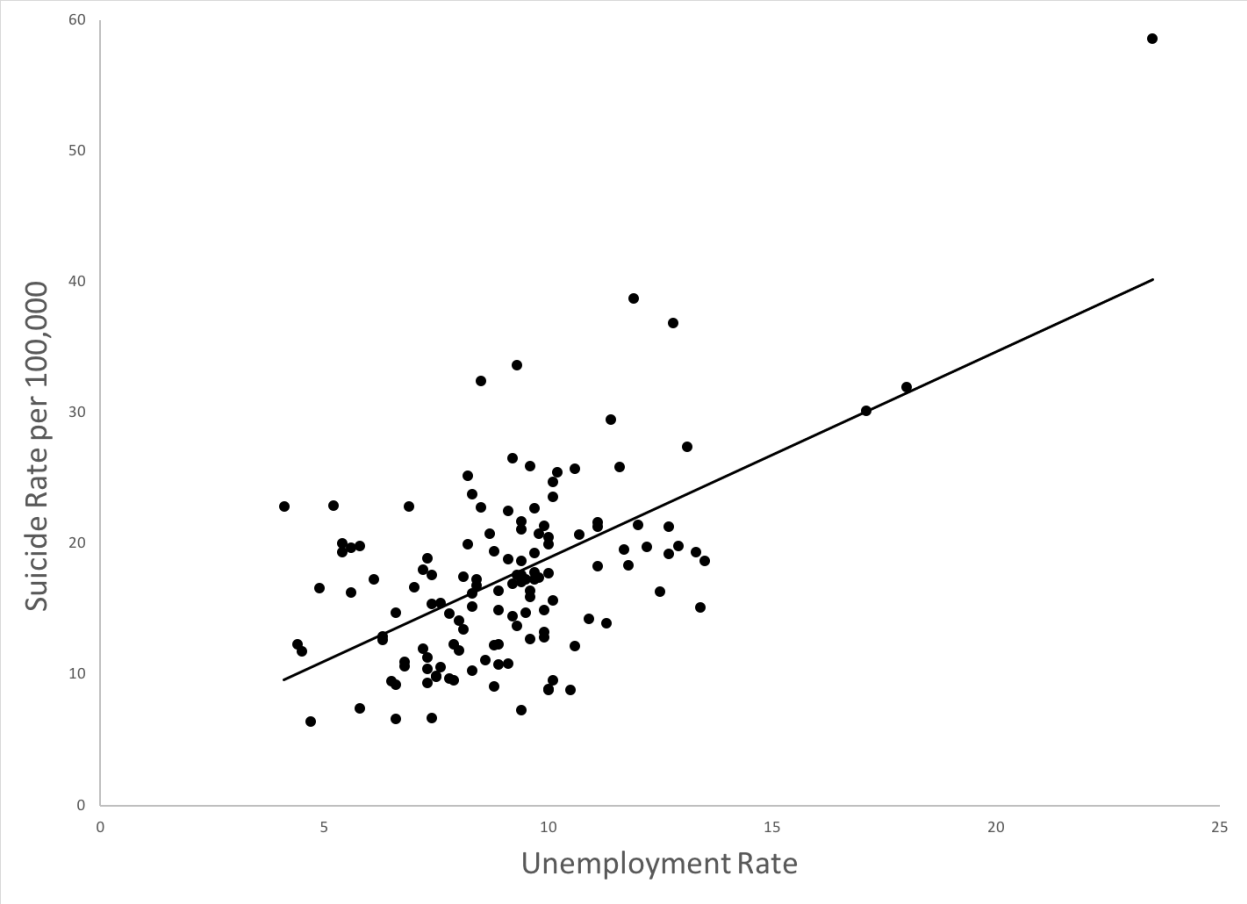


Figure 7. Scatterplot of the suicide rate and unemployment rate in Alberta, 2016



APPENDIX B: PREDICTORS

Alberta Health Care Insurance Plan Registry

Residency Flag (0/1)

Sex (0/1)

Age

Social Proxy: Registered First Nations (0/1)

Social Proxy: Income Support (0/1)

Social Proxy: Child Intervention (0/1)

Social Proxy: Other (0/1)

Local Geographic Area: Metropolitan (0/1)

Local Geographic Area: Metropolitan Influence (0/1)

Local Geographic Area: Urban (0/1)

Local Geographic Area: Urban Influence (0/1)

Local Geographic Area: Rural Centre (0/1)

Local Geographic Area: Rural (0/1)

Local Geographic Area: Rural Remote (0/1)

Latitude of Residential Postal Code

Longitude of Residential Postal Code

Physician Claims

Total Cost

Total Physician Services: General Practitioner

Total Physician Services: Psychiatrist

Total Physician Services: Other

Total Diagnoses, Category 1 (ICD9: 291* or 292* or 303* or 304* or (305* and not 305.1))

Total Diagnoses, Category 2 (ICD9: 295* or 301.2)

Total Diagnoses, Category 3 (ICD9: 296* or 298.0 or 300.4 or 301.1 or 309* or 311*)

Total Diagnoses, Category 4 (ICD9: 297* or (298* and not 298.0))

Total Diagnoses, Category 5 (ICD9: 308* or (300* and not 300.4))

Total Diagnoses, Category 6 (ICD9: 301* not 301.1 and not 301.2)

Total Diagnoses, Category 7 (ICD9: 302*)

Total Diagnoses, Category 8 (ICD9: 306* or 316*)

Total Diagnoses, Category 9 (ICD9: 307*)

Total Diagnoses, Category 10 (ICD9: 290* or 293* or 294* or 310*)

Total Diagnoses, Category 11 (ICD9: 299* or 312* or 313* or 314* or 315*)

Total Diagnoses, Category 12 (ICD9: 317* or 318* or 319*)

Total Diagnoses, Category 13 (ICD9: Other)

<u>Ambulatory Care</u>

Total Emergency Department Visits, Parasuicide Diagnosis, Triage Category 1

Total Emergency Department Visits, Parasuicide Diagnosis, Triage Category 2

Total Emergency Department Visits, Parasuicide Diagnosis, Triage Category 3

Total Emergency Department Visits, Parasuicide Diagnosis, Triage Category 4

Total Emergency Department Visits, Parasuicide Diagnosis, Triage Category 5

Total Emergency Department Visits, Parasuicide Diagnosis, Triage Category 6

Total Emergency Department Visits, Mental Health Diagnosis, Triage Category 1

Total Emergency Department Visits, Mental Health Diagnosis, Triage Category 2

Total Emergency Department Visits, Mental Health Diagnosis, Triage Category 3

Total Emergency Department Visits, Mental Health Diagnosis, Triage Category 4

Total Emergency Department Visits, Mental Health Diagnosis, Triage Category 5

Total Emergency Department Visits, Mental Health Diagnosis, Triage Category 6

Total Emergency Department Visits, Other Diagnosis, Triage Category 1

Total Emergency Department Visits, Other Diagnosis, Triage Category 2

Total Emergency Department Visits, Other, Diagnosis Triage Category 3

Total Emergency Department Visits, Other Diagnosis, Triage Category 4

Total Emergency Department Visits, Other Diagnosis, Triage Category 5

Total Emergency Department Visits, Other Diagnosis, Triage Category 6

Total Mental Health Department Ambulatory Care Visits, Parasuicide Diagnosis

Total Mental Health Department Ambulatory Care Visits, Mental Health Diagnosis

Total Mental Health Department Ambulatory Care Visits, Other Diagnosis

Total Other Facility Department Care Visits, Parasuicide Diagnosis

Total Other Facility Department Care Visits, Mental Health Diagnosis

Total Other Facility Department Care Visits, Other Diagnosis

Inpatient

Total Inpatient Days, Psychiatric

Total Inpatient Days, Maternal

Total Inpatient Days, Other

Pharmaceutical Information Network

Total Unique Drug Identification Numbers, Mental Health (ATC: N05* or N06*)

Total Drug Days, Mental Health (ATC: N05* or N06*)

Total Unique Drug Identification Numbers, Non-Mental Health

Total Drug Days, Non-Mental Health

Disease Registry (Quarter of Diagnosis Forward)

Affective Disorder (0/1)

Anorexia (0/1)

Anxiety Disorder (0/1)

Asthma (0/1)

Atrial Fibrillation (0/1)

Chronic Kidney Disease (0/1)

Chronic Obstructive Pulmonary Disorder (0/1)

Congestive Heart Failure (0/1)

Dementia (0/1)

Diabetes (0/1)

End-Stage Renal Disease (0/1)

Epilepsy (0/1)

Gout (0/1)

Guillain-Barré Syndrome (0/1)

Hypertension (0/1)

Inflammatory Bowel Disease (0/1)

Ischemic Heart Disease (0/1)

Liver Cirrhosis (0/1)

Lupus (0/1)

Motor Neuron Disease (0/1)

Multiple Sclerosis (0/1)

Non-Organic Psychosis (0/1)

Organic Psychosis (0/1)

Osteoarthritis (0/1)

Osteoporosis (0/1)

Parkinson's Disease (0/1)

Rheumatoid Arthritis (0/1)

Schizophrenia (0/1)

Shingles (0/1)

Sleep Apnea (0/1)

Stroke (0/1)

Substance Abuse (0/1)

Ecologic

Local Geographic Area: Suicide Rate

Local Geographic Area: Proportion Registered First Nations

Local Geographic Area: Proportion Income Support

Local Geographic Area: Proportion Child Intervention

Local Geographic Area: Proportion Other

APPENDIX C: LITERATURE REVIEW SUMMARY