Journal of Affective Disorders 264 (2020) 107–114

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

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?

Michael Sanderson[a,*], Andrew GM Bulloch[b], JianLi Wang[c], Tyler Williamson[d], Scott B Patten[e]

[a] Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, TRW, 4th Floor, Room 4D66, 3280 Hospital Drive NW, Calgary, Alberta, Canada
[b] Hotchkiss Brain Institute, Department of Psychiatry, Cumming School of Medicine, University of Calgary, TRW, 4th Floor, Room 4D67, 3280 Hospital Drive NW, Calgary, Alberta, Canada
[c] School of Epidemiology, Public Health and Preventive Medicine, Department of Psychiatry, Faculty of Medicine, University of Ottawa, Royal Ottawa Mental Health Centre, 1145 Carling Avenue, Ottawa, Ontario, Canada
[d] Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, TRW, 3rd Floor, Room 3D15, 3280 Hospital Drive NW, Calgary, Alberta, Canada
[e] Department of Community Health Sciences, Department of Psychiatry, Cumming School of Medicine, University of Calgary, TRW, 4th Floor, Room 4D66, 3280 Hospital Drive NW, Calgary, Alberta, Canada

ARTICLE INFO

Keywords:
suicide
prediction
machine learning
artificial intelligence
administrative data

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.

1. 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 (Alberta Vital Statistics 2019). 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 (Alberta Vital Statistics 2019). 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

* Corresponding author.
E-mail address: michael.sanderson@gov.ab.ca (M. Sanderson).

https://doi.org/10.1016/j.jad.2019.12.024
Received 14 November 2019; Received in revised form 11 December 2019; Accepted 13 December 2019
Available online 14 December 2019
0165-0327/

(Alberta Vital Statistics 2019). 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 (Pisani et al., 2016).

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 (Mulder et al., 2016; Large et al., 2016) and although statistical models have been developed that predicted suicide better than chance and better than clinicians (Chan et al., 2016; Huang et al., 2017; Carter et al., 2017; Ribeiro et al., 2016), 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 (Ribeiro et al., 2016).

In an earlier study (Sanderson et al., 2019) (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 was 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 (Pisani et al., 2016), 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 (The XGBoost Contributors 2019).

2. Model classes

2.1. Neural Networks

Neural networks are a flexible class of machine learning models that were inspired by neuroscience (Goodfellow et al., 2016). 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 (Goodfellow et al., 2016; Hornik, 1991), 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.

2.2. 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 (The XGBoost Contributors 2019). 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.

2.3. 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. Methods

3.1. 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 (Alberta Health 2017). 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 (Karmakar et al., 2016). 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).

3.2. 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.

3.3. 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 (Alberta Vital Statistics 2019). 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.

3.4. 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 (James et al., 2015). 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 (James et al., 2015). 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 (James et al., 2015). 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) (James et al., 2015).

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 (Hanley and McNeil, 1982), 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.

3.5. 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

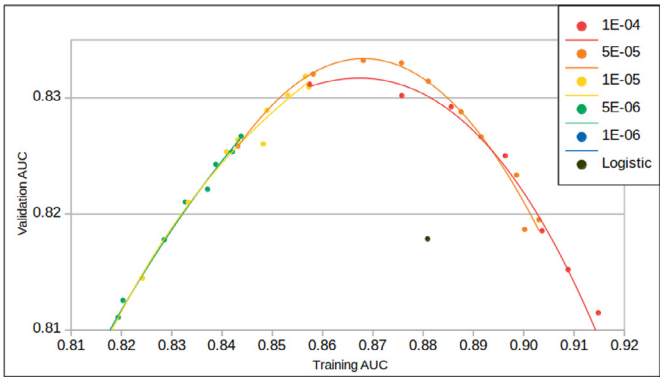

Fig. 1. Smoothed performance trajectories for the 8 neuron FNN model configurations and logistic regression from the log-FNN Study.

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, Fig. 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 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.

4. Results

4.1. 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).

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

4.2. 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 (CalibratedClassifierCV (2019)). 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 (Figs. 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.

4.3. 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: Figs. 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

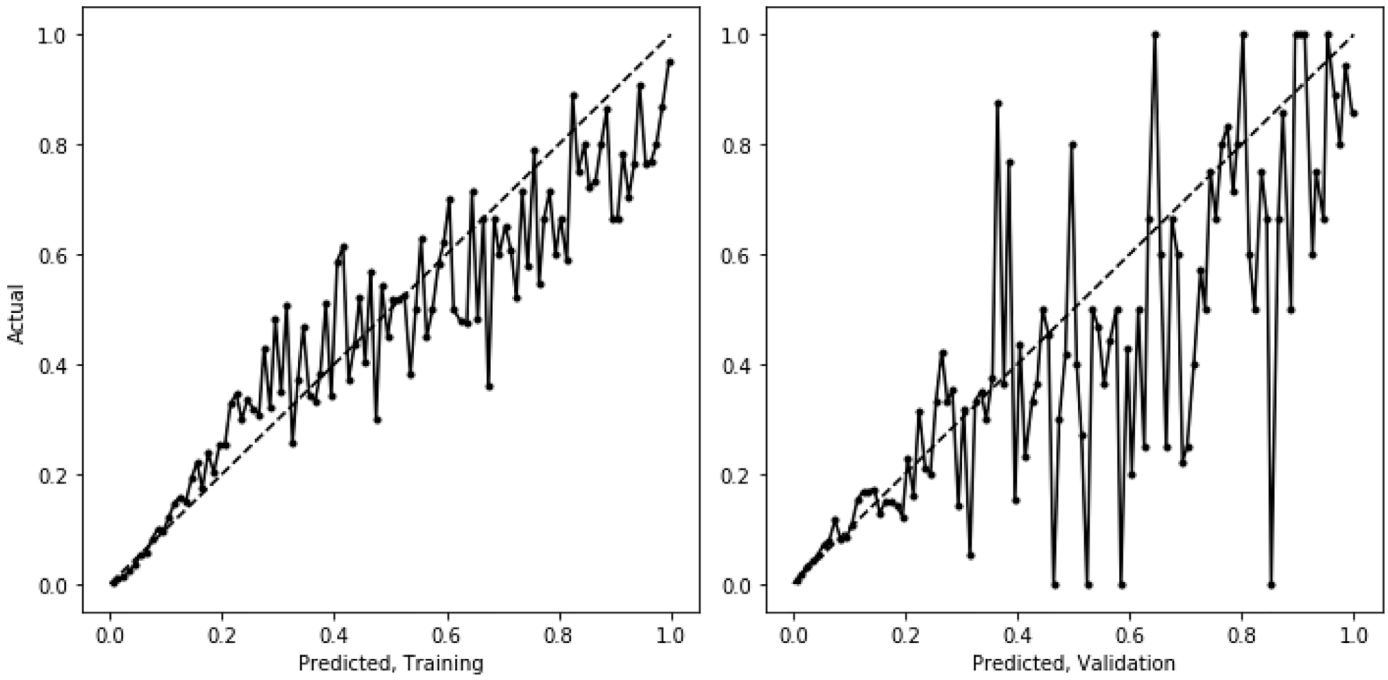

Fig. 2. Calibration curves, logistic regression, platt calibration.

also the case with the 1D-CNN and XGB model configurations but these figures are not included for the sake of brevity.

5. 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: Fig. 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

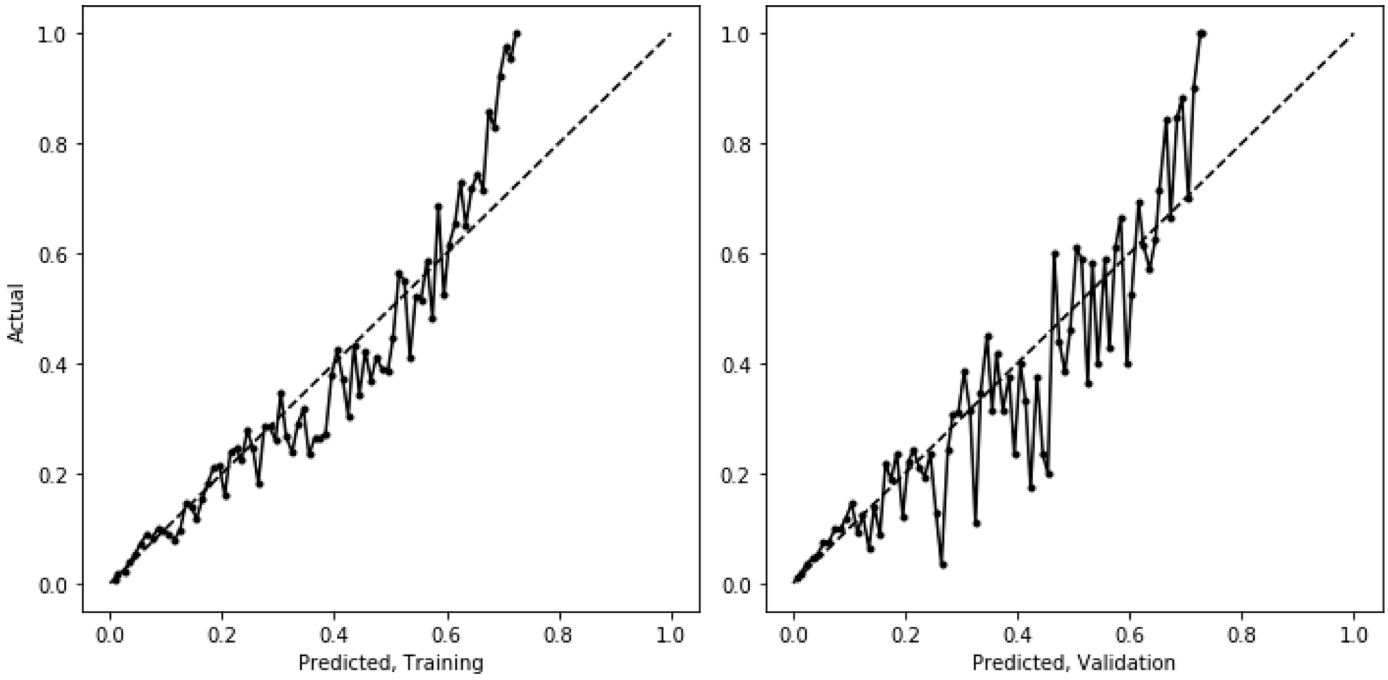

Fig. 3. Calibration curves, gradient boosted trees, platt calibration.

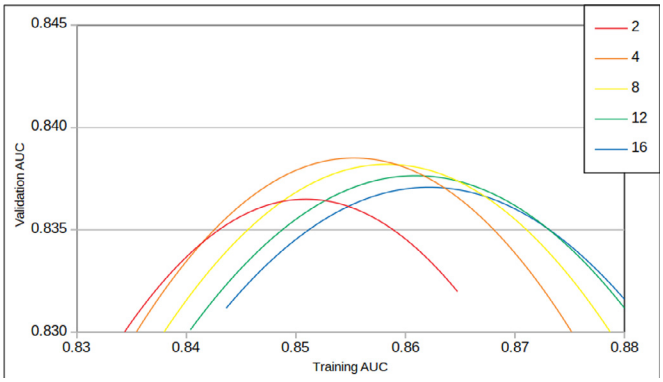

Fig. 4. Smoothed performance trajectories for the FNN model configurations, quarters.

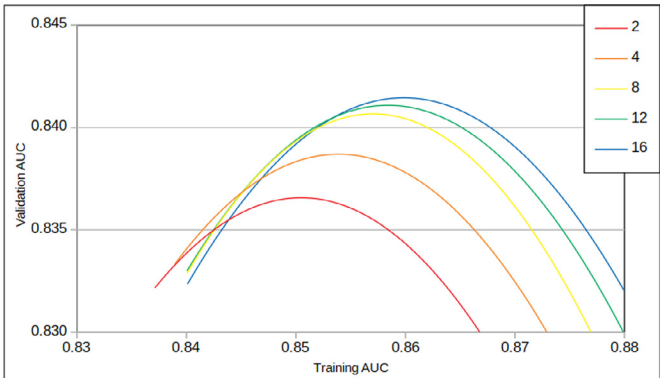

Fig. 5. Smoothed performance trajectories for the GRU model configurations, quarters.

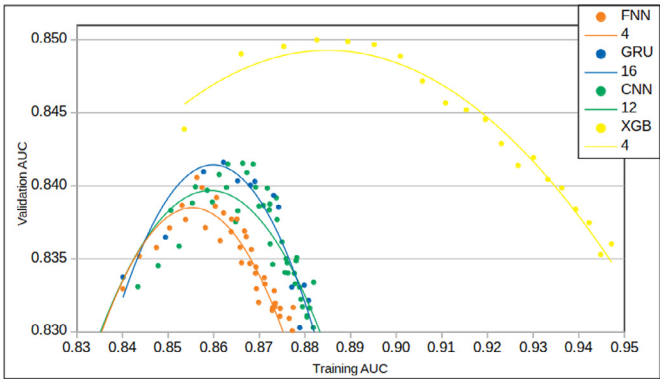

Fig. 6. Performance trajectories for the optimal FNN, GRU, CNN, XGB configurations, quarters.

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: Figs. 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: Fig. 6). The prevention framework suggested by Pisani et al., 2016, 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 (Tran et al., 2014; Pisani et al., 2016), 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 (Saunders et al., 2014; Katz et al., 2017; Chan et al., 2016; Carter et al., 2017; Large et al., 2016). 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.

6. 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.

Role of the funding source

No funding source.

Institutional review

This study was approved by the University of Calgary Conjoint Health Research Ethics Review Board.

CRediT authorship contribution statement

Michael Sanderson: Visualization, Data curation, Formal analysis, Writing - original draft. **Andrew GM Bulloch:** Visualization, Formal analysis, Writing - review & editing. **JianLi Wang:** Visualization, Formal analysis, Writing - review & editing. **Tyler Williamson:** Visualization, Formal analysis, Writing - review & editing. **Scott B Patten:** Visualization, Formal analysis, Writing - review & editing, Supervision.

Declaration of Competing Interest

None

Acknowledgements

None

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jad.2019.12.024.

Appendix A. Figures and Tables

Fig. 1–6 and Table 1

Appendix B. Predictors

Alberta Health Care Insurance Plan (AHCIP) 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

Supplemental Enhanced Service Event (SESE) Physician Service Payment 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)

Morbidity and Ambulatory Care Abstract Reporting (MACAR) Ambulatory Care Services

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

Morbidity and Ambulatory Care Abstract Reporting (MACAR) Inpatient Hospitalizations

Total Inpatient Days, Psychiatric
Total Inpatient Days, Maternal
Total Inpatient Days, Other

Pharmaceutical Information Network Community Pharmacy Dispense Records

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

References

https://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html 2019.
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.
Carter, G., Milner, A., McGill, K., Pirkis, J., Kapur, N., Spittal, M.J, 2017b. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. Br J Psychiatry.
Carter, G., Milner, A., McGill, K., Pirkis, J., Kapur, N., Spittal, M.J, 2017a. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. Br. J. Psychiatry.
Chan, M.K., Bhatti, H., Meader, N., Stockton, S., Evans, J., O'Connor, R.C., et al., 2016b. Predicting suicide following self-harm: systematic review of risk factors and risk scales. Br J Psychiatry 209 (4), 277–283.
Chan, M.K., Bhatti, H., Meader, N., Stockton, S., Evans, J., O'Connor, R.C., et al., 2016a. Predicting suicide following self-harm: systematic review of risk factors and risk scales. Br. J. Psychiatry 209 (4), 277–283.
Goodfellow, I., Bengio, Y., Courville, A, 2016. Deep Learning. MIT Press, pp. 192–195.
Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiving operating characteristic (ROC) curve. Radiology 143, 29–36.
Hornik, K., 1991. Approximation capabilities of multilayer feedforward networks. Neural Netw. 4 (2), 251–257 1991.
Huang, X., Ribiero, J.D., Musacchio, K.M., Franklin, J.C, 2017. Demographics as predictors of suicidal thoughts and behaviors: a meta-analysis. PLoS ONE 12 (7), e0180793.
James, G., Witten, D., Hastie, T., Tibshirani, R, 2015. An Introduction to Statistical Learning, With Applications in R, 6th Printing. Springer, New York.
Karmakar, C., Luo, W., Tran, T., Berk, M., Venkatesh, S, 2016. Predicting risk of suicide attempt using history of physical illnesses from electronic medical records. JMIR Ment. Health 3, 3.
Katz, C., Randall, J.R., Sareen, J., et al., 2017. Predicting suicide with the sad persons scale. Depress Anxiety 34 (9), 809–816.
Large, M., Kaneson, M., Myles, N., Myles, H., Gunaratne, P., Ryan, C., 2016b. Meta-analysis of longitudinal cohort studies of suicide risk assessment among psychiatric patients: heterogeneity in results and lack of improvement over time. PLoS ONE 11 (6), e0156322.
Large, M., Kaneson, M., Myles, N., Myles, H., Gunaratne, P., Ryan, C, 2016a. Meta-analysis of longitudinal cohort studies of suicide risk assessment among psychiatric patients: heterogeneity in results and lack of improvement over time. PLoS ONE 11 (6), e0156322.
Mulder, R., Newton-Howes, G., Coid, J.W, 2016. The futility of risk prediction in psychiatry. Br. J. Psychiatry 209, 271–272.
Pisani, A.R., Murrie, D.C., Silverman, M.M, 2016b. Risk stratification using data from electronic medical records better predicts suicide risks than clinician assessments. Academic Psychiatry 40, 623–629.
Pisani, A.R., Murrie, D.C., Silverman, M.M, 2016a. Reformulating suicide risk formulation: from prediction to prevention. Acad. Psychiatry 40 (4), 623–629 Aug.
Ribeiro, J.D., Franklin, J.C., Fox, K.R., Bentley, K.H., Kleiman, E.M., Chang, B.P., Nock, M.K, 2016b. Suicide as a complex classification problem: machine learning and related techniques can advance suicide prediction – a reply to Roaldset. Psychol. Med. 46, 2009–2010.
Ribeiro, J.D., Franklin, J.C., Fox, K.R., Bentley, K.H., Kleiman, E.M., Chang, B.P., Nock, M.K, 2016a. Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta- analysis of longitudinal studies. Psychol. Med. 46, 225–236.
Sanderson, M., Bulloch, A., Wang, J., Williamson, T., Patten, S, 2019. Predicting death by suicide using administrative health care system data: can feedforward neural network models improve upon logistic regression models? J. Affect. Disord. 257, 741–747.
Saunders, K., Brand, F., Lascelles, K., Hawton, K, 2014. The sad truth about the Sadpersons scale: an evaluation of its clinical utility in self-harm patients. Emerg. Med. J. 31 (10), 796–798.
The XGBoost Contributors. (2019) https://xgboost.readthedocs.io/en/latest/tutorials/model.html.
Tran, T., Luo, W., Phung, D., Harvey, R., Berk, M., Kennedy, R.L., Venkatesh, S, 2014. Risk stratification using data from electronic medical records better predicts suicide risks than clinician assessments. BMC Psychiatry 14, 76.