Machine Learning for Injuries Cause of Death Assignment: A New Method for the Global Burden of Disease Study

Kareha Agesa

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Public Health

University of Washington

2020

Committee:

Mohsen Naghavi

Abraham Flaxman

Program Authorized to Offer Degree:

Global Health

University of Washington

Abstract

Machine Learning for Injuries Cause of Death Assignment: A New Method for the Global Burden of Disease Study

Kareha Agesa

Chair of the Supervisory Committee:

Mohsen Naghavi, MD, MPH, PhD

Global Health

Globally, injuries were responsible for 8% of deaths in 2017 and have been a neglected source of burden, especially in many low income countries.[1,2] Effective public health interventions and decision making rely on accurate estimates of injuries burden; however, inconsistent and unreliable coding of injuries deaths has complicated this task over time. In particular, a large portion of injuries deaths are coded to Exposure to unspecified factor (International Classification of Diseases (ICD) 10: X59) and Unspecified event, undetermined intent (ICD 10: Y34), when there is insufficient information regarding the circumstances of an injuries death.[3] These garbage-coded deaths, or deaths assigned to ICD codes that are insufficiently specific or for which death is impossible, have a deleterious effect on cause-specific public health interventions.

The Global Burden of Disease Study (GBD) has developed an overall algorithm for redistributing garbage codes to a predefined cause list in order to attribute these deaths to more informative causes.[4] This process involves grouping similar ICD codes into “packages” and defining a list of “target causes” for each package to then be redistributed onto by age, sex, location, and year. Current redistribution methods are based on either statistical models, literature review, or expert opinion; however, a growing field of interest in the GBD is the use of multiple cause of death (MCoD) data to inform garbage code redistribution.[1] In GBD 2019, a novel regression method was introduced using MCoD data to redistribute X59 and Y34 deaths, however it relied largely on an algebra-based preliminary proportional redistribution method prior to modeling.[5] This analysis seeks to improve upon this preliminary method using machine learning.

Contents

Introduction

Timely and reliable cause of death data is important in public health planning and decision making.[4] Lack of cause of death data inhibits our understanding of what people are dying from and prohibits the formation of targeted interventions for these diseases. Reliable death data depends not only on a functioning vital registration system, but also consistent use of an epidemiological classification system such as the International Classification of Diseases (ICD). Though the ICD is considered the global health standard for reporting disease mortality and morbidity, global inconsistencies in completing death certificates and systematic underreporting of certain causes of death (such as suicide or HIV/AIDS) often lead to unreliable and inconsistent data.[6] Moreover, despite the ICD having a defined set of rules for determining the underlying cause of death, a large proportion of global deaths are assigned to garbage codes, codes that either are not useful for public health analysis or for which death is impossible.[7,8]

The Global Burden of Disease (GBD) Study has developed methods to redistribute garbage coded deaths to a predefined list of 286 causes of death estimated in the study. Similar garbage codes are grouped into "redistribution packages" and a group of target causes are identified for each package based on pathophysiology and disease classification. These packages are then redistributed to target causes using one of three methods: proportional redistribution; regression analyses, or literature review supplemented with an expert-based algorithm.[4] Of the cause of death data in the GBD, certain garbage codes have historically comprised a large proportion of deaths, such as X59 (ICD 10: "Exposure to unspecified factor") and Y34 (ICD 10: "Unspecified event, undetermined intent") for injuries deaths.[1] In past cycles of the GBD, these codes have been redistributed using an expert-based algorithm, however in GBD 2019, a novel regression method was introduced using multiple cause of death (MCoD) data.[5] This analysis seeks to build upon the GBD 2019 methods developed for redistributing all codes encompassing the X59 and Y34 garbage code packages using 4 different machine learning classifiers: naïve Bayes, random forest, gradient boosted trees (GBT), and deep neural networks (DNN).

Related Work

The use of supervised machine learning methods for classifying cause of death has been well documented, with many algorithms showing promise in assigning cause of death for verbal autopsy (VA) data. Flaxman et.al showed that a random forest method outperformed physician-

certified verbal autopsy in assigning cause of death at both the individual and population level, using the chance-corrected concordance and cause-specific mortality fraction accuracy.[9] Additionally, Mujtaba et.al found random forest models parameterized using expert-driven feature selection achieved the highest overall accuracy, macro precision, and macro F-measure in automatic classification of ICD-10 related cause of death using autopsy reports.[10] While simple, naïve Bayes classifiers have also been shown to outperform other cause of death classifiers such as InterVA-4 and an open source version of the Tariff method in terms of sensitivity, specificity, and the cause-specific mortality fraction accuracy.[11]

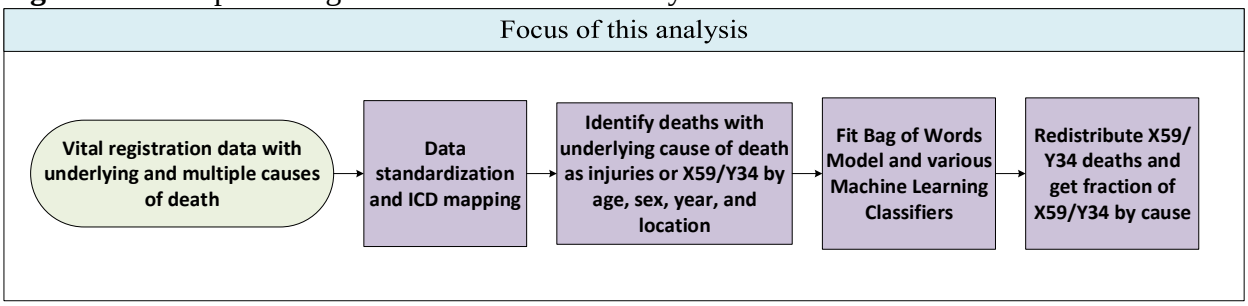
Despite their recent popularity and known improvements over random forest models, GBT are not widely used in classifying causes of death, though they have shown promise in predicting suicide deaths from health administrative data.[12] Neural networks, though popular in business applications and in image classification and, are also not widely used in classifying causes of death. However, recently, Jeblee et.al used a feed-forward network to classify cause of death from VA narratives achieving a high sensitivity for individual cause of death assignment.[13]

Novelty

Utilizing the promising nature of machine learning algorithms in classifying cause of death, this analysis seeks to improve upon methods of X59 and Y34 garbage code redistribution in the GBD. The current redistribution process involves a multi-step approach with a modeling framework similar to that of intermediate causes such as sepsis.[5] In this framework, MCoD data was first used to identify deaths with an underlying cause of death as either a GBD injuries cause, or an ICD code in the X59 or Y34 packages (**Appendix Table 1**). A cause-specific redistribution proportion was then derived using the pattern of Nature of Injury codes in the causal chain of the MCoD data and the probability of a death being coded to X59/Y34 or a GBD injuries cause. These proportions were then used to redistribute all MCoD deaths with the underlying cause of death as either X59 or Y34, and the fraction of X59 and Y34 deaths by age, sex, year, location, and injuries cause were then calculated and modeled using a mixed effects linear regression. These fractions were multiplied by GBD 2017 injuries cause of death estimates and the fraction of X59 and Y34 attributable to each cause was then used to redistribute X59 and Y34 deaths for all available cause of death data in the GBD.[5] This analysis will focus on improving the first step, involving redistributing X59 and Y34 deaths in the MCoD data (**Figure**

1). The goal of this analysis is to introduce for the first time machine learning algorithms into GBD garbage code redistribution and compare how this change affects redistribution results in the input data.

Figure 1. Conceptual diagram of redistribution analysis



Methods

Data

This analysis used all nationally-representative, individual-level MCoD data available in the GBD. This included 513 location-years of data from Brazil, 20 location-years of data from Columbia, 13 location-years of data from Italy, 256 location-years of data from Mexico, 10 location-years from Taiwan, and 1,938 location-years of data from the United States. Data from Brazil, Colombia, Italy, Mexico, and Taiwan were ICD-10 coded and data from the United States contained both ICD-9 and ICD-10 coded deaths (**Table 1**). Data from Brazil, Mexico, and the United States were used at the subnational level, with the Colombia, Italy, and Taiwan data at the national level. No primary data were collected for this study, and all data were completely anonymous.

Demographic information such as age, sex, year, and location were extracted from each record, along with all ICD-coded death information. Deaths where an injuries-related ICD code was the underlying cause of death were each mapped to a most-detailed GBD injuries causes. Deaths that were not injuries related were dropped. Of the 104 million deaths available in these records, 10% were injuries related, with 33% of these injuries deaths being garbage coded (**Table 1**). Of the injuries garbage coded deaths 15% were X59 and 21% were Y34, though this fraction varied greatly by country (**Figure 2**).

Table 1. Data sources used for estimation

Location	Data Source	Years	ICD Classification	Individual Records	Injuries Records
Brazil	Mortality Information System (SIM)	1999 – 2017	ICD 10	16,932,859	2,387,101
Colombia	National Administrative Department of Statistics (DANE)	1998 – 2017	ICD 10	3,624,771	699,070
Italy	Italian National Institute of Statistics (ISTAT)	2003 – 2015	ICD 10	7,640,383	336,596
Mexico	National Institute of Statistics and Geography (INEGI)	2009 – 2016	ICD 10	4,336,713	551,998
Taiwan (province of China)	Cause of Death Data Statistics and Management System (CDDSM)	2008 – 2017	ICD 10	1,560,194	116,627
United States	National Vital Statistics System (NVSS)	1980 – 2017	ICD 9, ICD 10	70,344,838	6,466,471

Figure 2. Percent of injuries garbage coded deaths coded to **(a)** X59 and **(b)** Y34 by location.

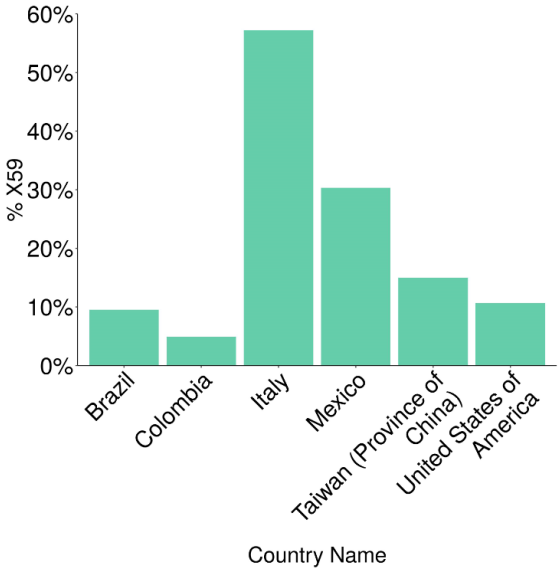


Figure 2a. Percent of injuries garbage coded deaths that are X59

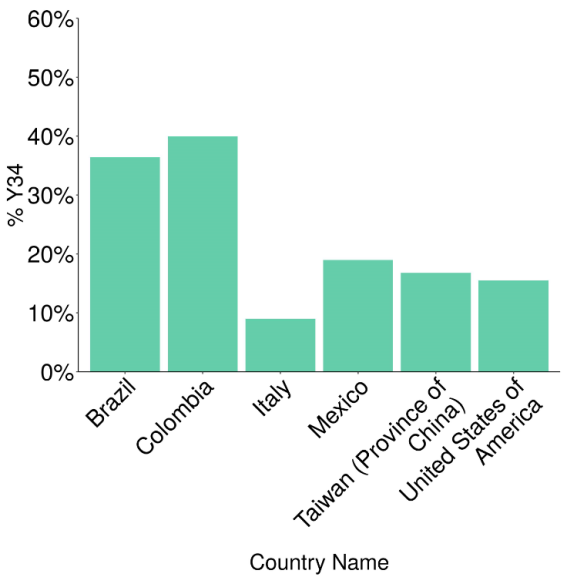


Figure 2b. Percent of injuries garbage coded deaths that are Y34

X59 and Y34 Packages

The list of target codes for which the X59 and Y34 packages can be redistributed onto are defined by the most-detailed injuries causes in the GBD. Since X59 is defined by unintentional injuries, it cannot be redistributed onto causes related to self-harm and interpersonal violence. Likewise, Exposure to forces of nature and Conflict and terrorism were dropped as targets for both packages. **Appendix Table 1** gives the ICD codes that define the X59 and Y34 packages along with the ICD codes that define the targets.

Machine Learning Classifiers

A bag of words model was used for feature extraction and to train subsequent predictive models separately for X59 and Y34. Data was split into 75% train, 25% test datasets, and binary features were created to identify ICD codes in the chain, along with demographic information such as age, sex, year, and location using the CountVectorizer in Python's scikit-learn module. This final feature matrix was used to fit a naïve Bayes, random forest, GBT, and DNN. Grid search was used to test varying model parameters, and 5-fold cross validation was performed with precision, sensitivity, accuracy, chance-corrected concordance (CCC), and chance-corrected cause-specific mortality fraction accuracy (CCCSMFA) as evaluation metrics. CCC was used to determine the best parameters for each classifier. Descriptions of each classifier and their chosen parameters are discussed below.

Naïve Bayes

Naïve Bayes is a popular machine learning algorithm that is considered simple, yet efficient. It is derived from Bayes' theorem and applies the strong independence assumption among all features.[14] For a class variable y and dependent feature vector x_i through x_n, Bayes' theorem states their relationship as[15]:

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n|y)}{P(x_1, \dots, x_n)} \quad (1)$$

Where the naïve independence assumption implies[15]:

$$P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y) \quad (2)$$

Which simplifies to the following for all i[15]:

$$P(y|x_1, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, \ldots, x_n)} \quad (3)$$

Several naïve Bayes classifiers that differ in their assumptions made regarding the distribution of $P(x_i|y)$ are available in Python's scikit-learn module. Multinomial, Bernoulli, and Complement Naïve Bayes were tried in this analysis and Multinomial Naïve Bayes was chosen based on its out of sample performance. This classifier assumes a multinomial distribution of $P(x_i|y)$ that is based on relative frequency counting. This distribution is parameterized by a smoothing parameter α which prevents zero probabilities for features not present in the training sample[15]. A grid search was used to determine the optimal value of α (**Table 2**).

Random Forest

Random forest is an ensemble machine learning algorithm made up of a collection of decision trees. Decision trees are flowchart like structures that use recursive binary splitting to split the input data using its features.[16] For classification problems, either the Gini Index or cross-entropy can be applied as the cost function used to determine the feature's split points. Both were tried and the Gini Index was used in this analysis based on results from the grid search. The formula for the Gini Index is shown[17]:

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2 \quad (4)$$

Where p_i is the probability of being classified to a given class i.

For classification problems, random forest models arrive at a final prediction by aggregating the predictions from the individual trees. These models often outperform single decision trees for a multitude of reasons. Random forests combat issues of overfitting common to decision trees through their use bagging and random feature selection to create an uncorrelated forest of trees.[16] In this analysis, grid search was used to tune values related to the number of trees, the maximum depth of each tree, and the number of features to use when looking for the best split. Chosen parameter values are shown in **Table 2.**

Gradient Boosted Trees

GBT are another ensemble method that utilizes a boosting method to sequentially builds trees, instead of combining the predictions from individual trees in parallel, like a random forest. A loss function is used to determine which trees are "weak learners", and sequential trees are

built and modified based on the errors of previous trees.[18] Simply stated, for x features and y class, the algorithm will fit a decision tree to the data[19]:

$$F_1(x) = y \quad (5)$$

Then fit the next decision tree to the residuals of the previous tree[19]:

$$h_1(x) = y - F_1(x) \quad (6)$$

And add this new tree to the algorithm[19]:

$$F_2(x) = F_1(x) + h_1(x) \quad (7)$$

The predictions of the final ensemble come from a weighted sum of the predictions made by previous trees. To prevent overfitting and reduce the contribution of any one tree, a constant learning rate is applied to scale the contribution of each weak learner.[18] Grid search was used to determine the optimal values of the learning rate, the number of trees, the maximum depth for each tree, along with other parameters related to each tree. All best parameters values for the X59 and Y34 models are given in **Table 2.**

Deep Neural Networks

The neural network is modeled based loosely off of the human brain, and is characterized by multiple layers of interconnected nodes used to transmit information. DNN consist not only of an input and an output layer, but also multiple hidden layers that process and transmit inputs. As inputs are fed forward through each layer, they are multiplied by a weight and a bias is added.[20] For an initial layer, this is actualized through this equation[20]:

$$y_j = f(\sum_i w_{i,j} a_i + b_i) \quad (8)$$

Where w_i are weights that are multiplied by the first layer nodes a_i and summed with a bias term b_i. These values are summed and passed to the activation function f then passed to output node y_j. In this analysis, a rectified linear unit (ReLU) activation function was chosen for hidden layers, and a softmax activation function was chosen for the output layer. Backpropagation is used to train and update the weights using an optimizer. Adam, a commonly used optimizer, was chosen as the optimizer in this analysis. The network in this analysis consisted of two hidden layers, and grid search was used to determine the number of epochs (the number of times the entire dataset is passed forward and backward through the network), the

batch size (the number of training samples passed through), and the number of nodes in each hidden layer (**Table 2**).

Table 2. Best parameters for each classifier for X59 and Y34 models

Classifier	**Parameters**	
	X59 model	Y34 model
Naïve Bayes	Alpha = 0.05	Alpha = 0.05
Random Forest	Number of trees = 170 Max depth = 750 Max Features = Square root of features Criterion = Gini	Number of trees = 160 Max depth = 600 Max Features = Square root of features Criterion = Gini
GBT	Number of trees = 130 Max Depth = 30 Learning Rate = 0.35 Minimum loss required to split = 0.4 Fraction of observations in each tree = 0.8	Number of trees = 130 Max Depth = 30 Learning Rate = 0.45 Minimum loss required to split = 0.8 Fraction of observations in each tree = 0.8
DNN	Epochs = 6 Batch Size = 30,000 Hidden Layers = 2 # Nodes 1st hidden layer = 8,000 # Nodes 2nd hidden layer = 1,000	Epochs = 5 Batch Size = 40,000 Hidden Layers = 2 # Nodes 1st hidden layer = 8,000 # Nodes 2nd hidden layer = 1,000

Model Evaluation

After splitting the observed data into 75%/25% training and test datasets, the test dataset was used as the basis for model evaluation. Given that this split was done using random sampling without replacement, it is possible that the cause distribution of the test data is reflective of that of all observed data. Additionally, because metrics such as the CCCSMFA are a function of the cause specific mortality fractions in the data, there is a potential for issues of data leakage.[21] To prevent this, a series of 500 test datasets were generated and used for model evaluation (**Figure 3**).[21] The cause distribution of each dataset was generated by randomly drawing the cause composition from a Dirichlet distribution, each time with alpha informed by the cause distribution in the actual test data. Rows of the test data, containing both ICD and demographic information, were randomly sampled with replacement in accordance to the fractions drawn from the Dirichlet distribution to generate each test dataset. Predictions for each test dataset came from the best chosen parameters for each classifier from the grid search (**Table 2**). Model performance for each generated test dataset was evaluated using precision, sensitivity, accuracy, CCC, and CCCSMFA, and overall descriptive statistics for each metric over the range of datasets

are given in **Table 3** and **Appendix Table 2**. The best model was chosen based on the CCC and refit on all observed data prior to generating predictions for the unobserved data.

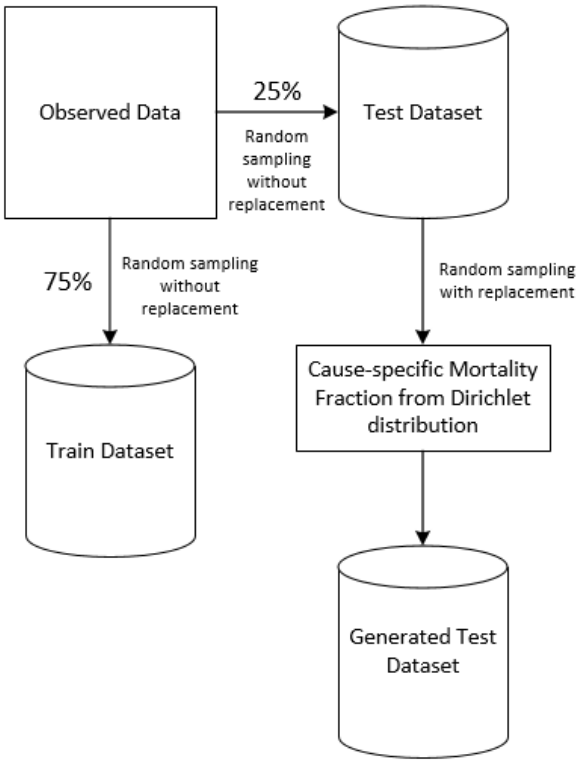


Figure 3. Process of creating training dataset and generating 500 test datasets.[22]

Results

Table 3 shows the mean precision, sensitivity, accuracy, CCC, and CCCSMFA for each of the classifiers for the X59 (**Table 3a**) and Y34 (**Table 3b**) models. For X59, overall the DNN performed best in terms of mean CCC (0.7998), precision (0.6059), and accuracy (0.8486), however GBT showed slightly better performance than the DNN in terms of mean CCCSMFA (0.8138), and the random forest classifier performed best in terms of mean sensitivity (0.6382) (**Table 3a**). For the Y34 model, the DNN performed the best in terms of mean CCC (0.8060), CCCSMFA (0.8573), precision (0.6113), and accuracy (0.8833); however similarly to the X59 model, the random forest classifier performed best in terms of mean sensitivity (0.6521) (**Table 3b**).

Table 3. Mean CCC, CCCSMFA, macro-precision, accuracy, and macro-sensitivity across the 500 test datasets for the X59 **(a)** and Y34 **(b)** models.

Classifier	**Mean CCC**	**Mean CCCSMFA**	**Mean Precision**	**Mean Accuracy**	**Mean Sensitivity**
Naïve Bayes	0.7803	0.7452	0.5420	0.8042	0.6234
Random Forest	0.7589	0.8018	0.5535	0.8263	**0.6382**
GBT	0.7997	**0.8138**	0.5890	0.8443	0.6079
DNN	**0.7998**	0.8109	**0.6059**	**0.8486**	0.6380

Table 3a. Results for X59

Classifier	**Mean CCC**	**Mean CCCSMFA**	**Mean Precision**	**Mean Accuracy**	**Mean Sensitivity**
Naïve Bayes	0.7503	0.7637	0.5576	0.8428	0.6072
Random Forest	0.7672	0.8344	0.5740	0.8680	**0.6521**
GBT	0.8026	0.8465	0.5948	0.8825	0.6316
DNN	**0.8060**	**0.8573**	**0.6113**	**0.8833**	0.6483

Table 3b. Results for Y34

In terms of final cause of death assignments, the DNN was chosen as the best classifier for both X59 and Y34 based on its mean out of sample CCC. The top ten GBD injuries causes with the highest proportion of redistributed X59 and Y34 deaths in the DNN vs GBD 2019 are shown in **Figure 4**. While the results from GBD 2019 had a large proportion of X59 deaths being classified as falls (74%), the majority of deaths from the DNN were redistributed to motor vehicle road injuries (43%) (**Figure 4a**). Likewise, the DNN predicted a higher proportion of X59 deaths to be redistributed to pedestrian road injuries (23%) and adverse effects of medical treatment (16%), than was used in GBD 2019 (7% and 1%, respectively) (**Figure 4a**). For the Y34 model, though physical violence by firearm received the largest proportion of deaths in GBD 2019 (20%), adverse effects of medical treatment was predicted to receive the highest by the DNN (21%) (**Figure 4b**). Likewise, the DNN predicted a much higher proportion of deaths to be redistributed to motor vehicle road injuries (18%), and much lower proportions for self-harm by other specified means (9%), physical violence by firearm (6%), and falls (6%) than was used in GBD 2019 (5%, 20%, 20%, and 14%, respectively) (**Figure 4b**).

Figure 4. Redistribution fractions by GBD injuries cause in GBD 2019 vs the DNN in the X59 **(a)** and Y34 **(b)** models

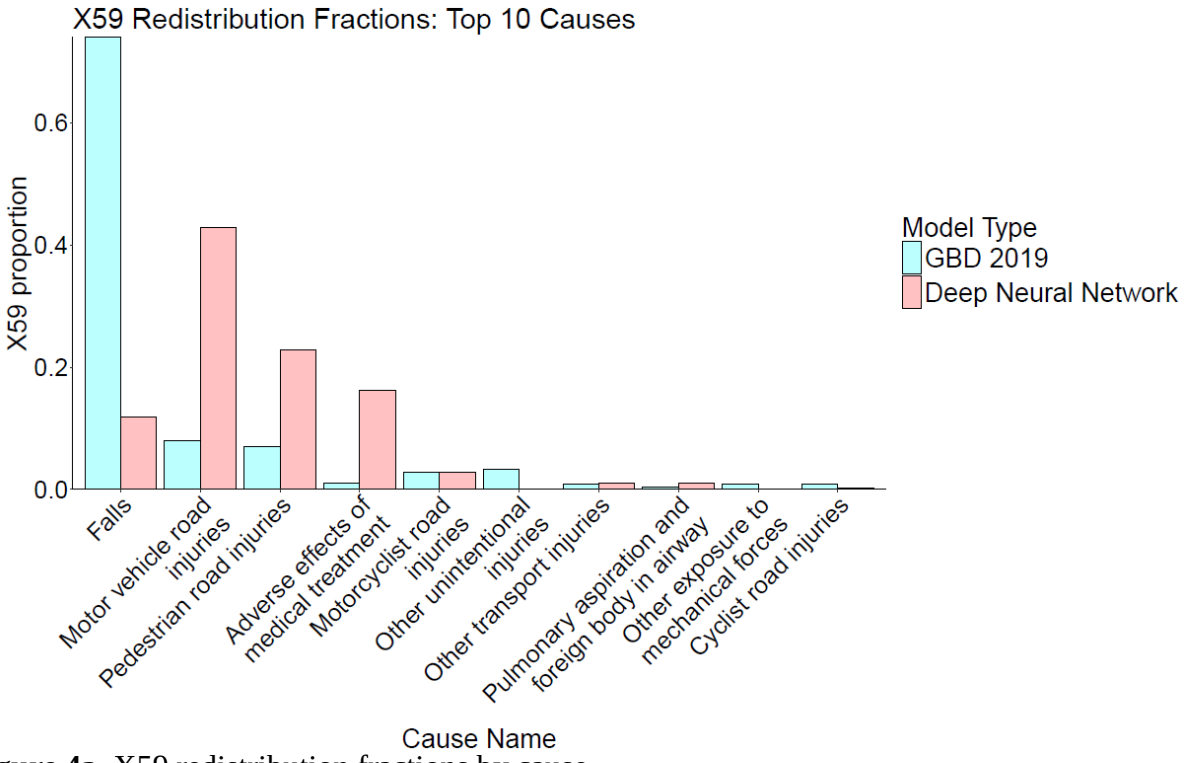


Figure 4a. X59 redistribution fractions by cause.

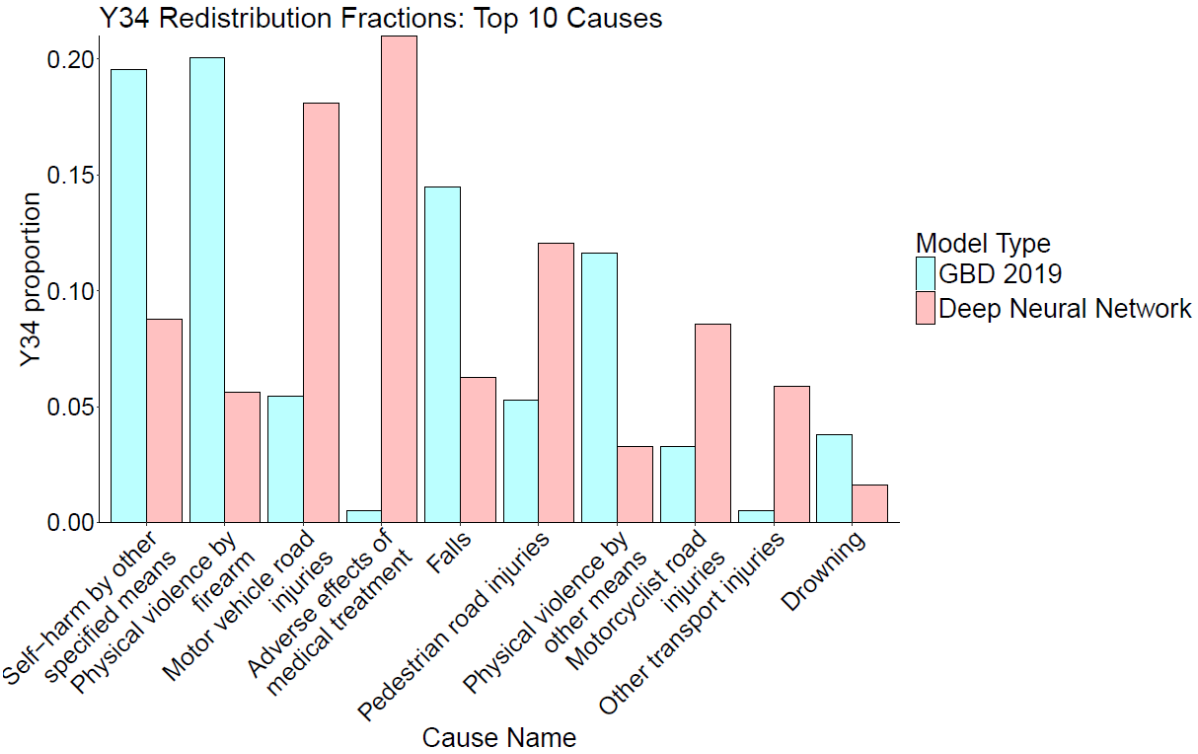


Figure 4b. Y34 redistribution fractions by cause

Overall, the 5 causes with the highest proportion of redistributed X59 deaths were motor vehicle road injuries (43%), pedestrian road injuries (23%), adverse effects of medical treatment (16%), falls (12%), and motorcyclist road injuries (3%) (**Figure 4a**). For Y34, the top 5 causes were adverse effects of medical treatment (21%), motor vehicle road injuries (18%), pedestrian road injuries (12%), self-harm by other specified means (9%), and motorcyclist road injuries (9%) (**Figure 4b**). In looking by country for X59, the trend in motor vehicle road injuries was largely driven by the United States (67% of US X59 deaths), while falls was driven largely by Mexico (44% of Mexico X59 deaths) (**Figure 5a**). For Y34, Italy drove the trend in adverse effects of medical treatment (61% of Italy Y34 deaths), while a large portion of the motorcyclist road injuries numbers were driven by Taiwan (32% of Taiwan Y34 deaths) (**Figure 5b**).

Figure 5. Top 5 causes **(a)** X59 and **(b)** Y34 redistributed onto by country

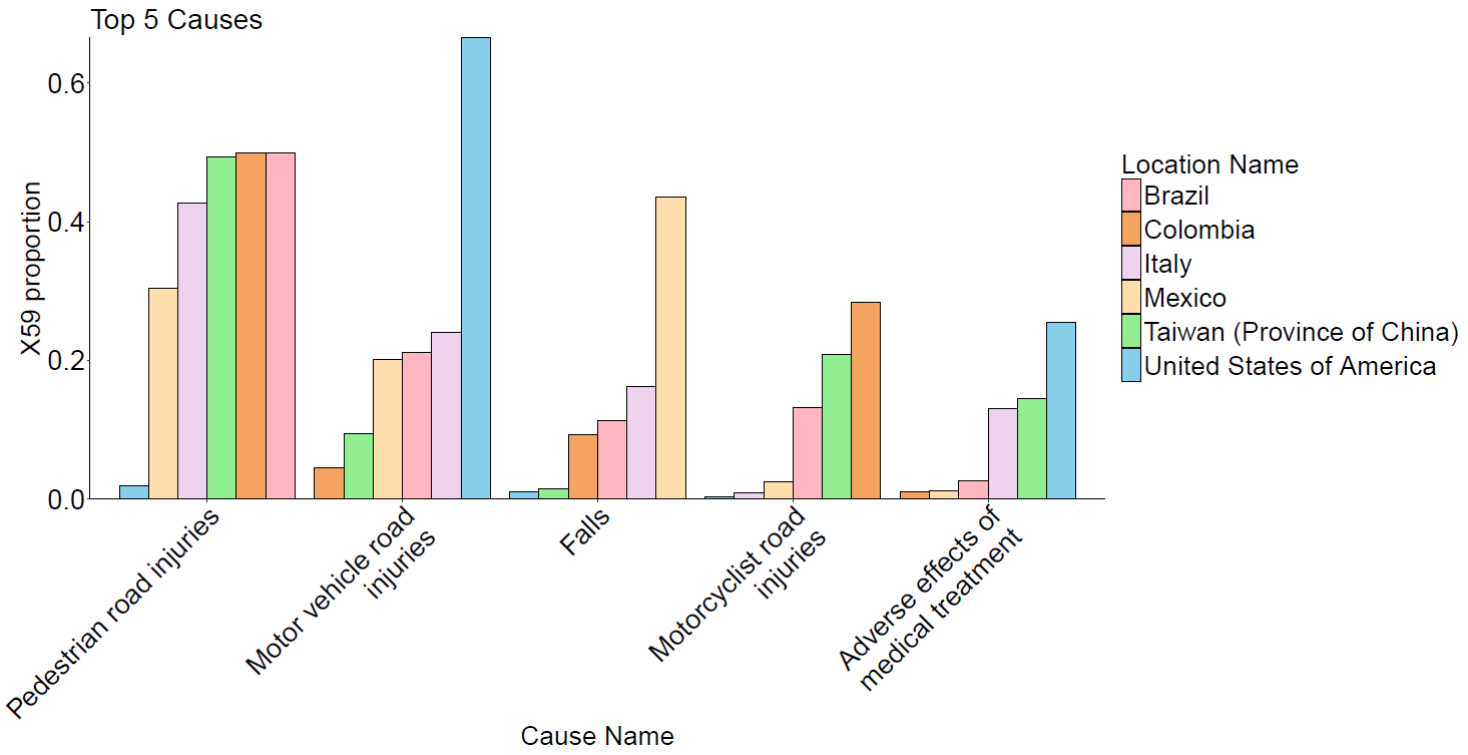


Figure 5a. Top 5 causes for X59

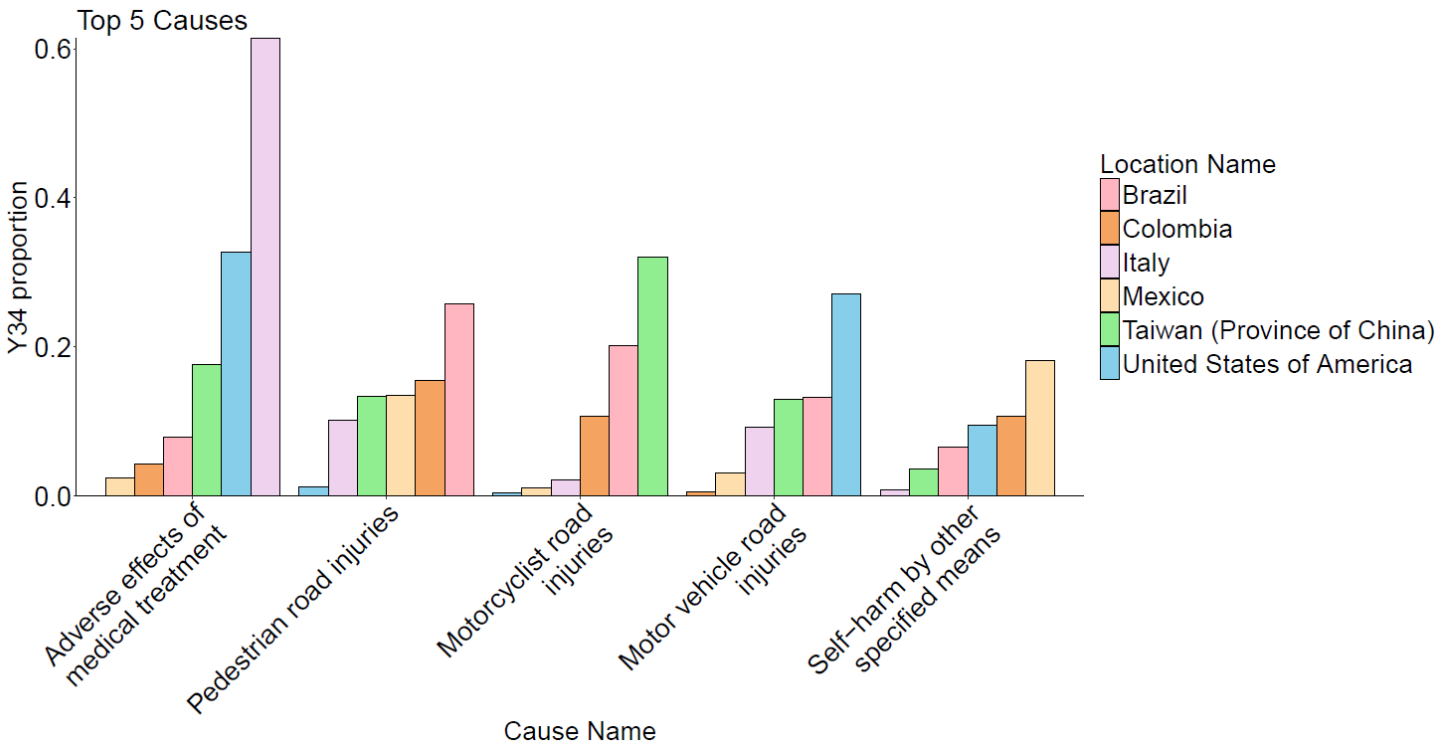


Figure 5b. Top 5 causes for Y34

Discussion

This analysis explored the use of various machine learning classifiers (naïve Bayes, random forest, GBT, and DNN) for individual cause of death assignment of unspecified injuries deaths using multiple cause of death vital registration data. Focusing on the CCC as the metric to evaluate out of sample model performance, the DNN was chosen as the best model for both the X59 and the Y34 models (mean CCC: 0.7998 and 0.8060 respectively); however, the GBT model also showed promising results (mean CCC: 0.7997 and 0.8026 respectively). In looking at evaluation metrics such as mean CCCSMFA, GBT showed slightly better results in the X59 model (0.8138) compared to the DNN (0.8109), but the DNN remained best for Y34 (0.8573). Random forest performed best for both X59 and Y34 in terms of mean sensitivity (0.6382 and 0.6521 respectively).

There are many differences in the redistribution proportions comparing the results of this analysis to the results from GBD 2019. Of particular interest might be the large decrease in deaths assigned to falls for X59 and the large increase in deaths assigned to adverse effects of medical treatment for Y34 in the DNN. There are several reasons that could have caused these changes. First, in GBD 2019, all Nature of Injury codes found in the chain of the MCoD data were grouped into custom categories based on the type of ICD code prior to performing any

analyses. However, in this analysis, all ICD coded information was used at the most-detailed level (though a sub analysis that also incorporated the hierarchical nature of the ICD was performed and yielded similar results as that above). This difference largely affects both the number and the nature of the features fed into each classifier in this analysis, and more work is necessary to determine which method is preferred. Additionally, the redistribution numbers created in GBD 2019 were a straightforward implementation yielded from the proportions in the input data.[5] For the DNN however, model behavior is slightly more complicated to understand and predict.[23] More work also needs to be done to understand which features are important in generating predictions.

Though this analysis was a great improvement in existing methods to redistribute X59 and Y34 related deaths in the GBD, there are several limitations to this study. First, while this study is only the first step in generating X59/Y34-related redistribution proportions for the GBD, it serves as the basis for later modeled results. Therefore, the lack of geographical diversity in the data sources (particularly from Sub-Saharan Africa) will have tremendous implications on the global interpretability of these results. Likewise, as the results from machine learning algorithms are highly dependent upon the input features, it could be expected that incorporating more data sources will result in large changes in results.[24]

Additionally, an abundance of X59/Y34 deaths in the input data could also lead to biased results. For example, in countries with a large number of deaths coded to X59, there could resultantly be a scarcity of deaths coded to GBD injuries causes, decreasing the amount of data the model can learn from. Moreover, it has been recorded that specificity in filling out a death certificate varies widely by country and age, with certain countries such as Sweden and Australia more likely to categorize chain causes such as femur fracture in older ages as an X59 death.[3] This could potentially lead to an age-specific lack of deaths coded to GBD injuries causes that also have a femur fracture in the chain. For this reason, it may be good to incorporate location and age specific redistribution priors in subsequent analyses.

This analysis introduced using machine learning classifiers as a novel method of garbage code redistribution to the GBD. This work had many strengths, including its abundance of nationally-representative individual-level vital registration MCoD data that allowed for detailed analyses by age, sex, location, and year. To further understand the implications of these results, more work should be done to track the implications of this analysis on unspecified injuries final

redistribution proportions in the GBD. However, this analysis was a great first step in elucidating trends in injuries related deaths.

References

1. Roth GA, Abate D, Abate KH, et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*. 2018;392(10159):1736-1788. doi:10.1016/S0140-6736(18)32203-7

2. Jamison DT, Breman JG, Measham AR, et al., eds. *Disease Control Priorities in Developing Countries*. 2nd ed. World Bank; 2006. Accessed February 12, 2020. http://www.ncbi.nlm.nih.gov/books/NBK11728/

3. Lu TH, Walker S, Anderson RN, McKenzie K, Bjorkenstam C, Hou WH. Proportion of injury deaths with unspecified external cause codes: a comparison of Australia, Sweden, Taiwan and the US. *Injury Prevention*. 2007;13(4):276-281. doi:10.1136/ip.2006.012930

4. Naghavi M, Makela S, Foreman K, O'Brien J, Pourmalek F, Lozano R. Algorithms for enhancing public health utility of national causes-of-death data. *Population Health Metrics*. 2010;8(1). doi:10.1186/1478-7954-8-9

5. GBD 2019 Diseases and Injuries, and Impairments Collaborators. Global burden of 369 diseases and injuries, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*. Published online Manuscript submitted for publication.

6. Mathers CD, Fat DM, Inoue M, Rao C, Lopez AD. Counting the dead and what they died from: an assessment of the global status of cause of death data. *Bull World Health Organ*. 2005;83(3):171-177.

7. World Health Organization, ed. *International Statistical Classification of Diseases and Related Health Problems. Vol. 3: Alphabetical Index*. 10. rev.; 2008.

8. Murray CJL, ed. *The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020 ; Summary*. Harvard School of Public Health [u.a.]; 1996.

9. Flaxman AD, Vahdatpour A, Green S, James SL, Murray CJ, Population Health Metrics Research Consortium (PHMRC). Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Popul Health Metr*. 2011;9:29. doi:10.1186/1478-7954-9-29

10. Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K, Al-Garadi MA. Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. Zhang Y, ed. *PLoS ONE*. 2017;12(2):e0170242. doi:10.1371/journal.pone.0170242

11. Miasnikof P, Giannakeas V, Gomes M, et al. Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. *BMC Medicine*. 2015;13(1). Accessed February 2, 2020. http://link.gale.com/apps/doc/A469138200/HWRC?u=wash_main&sid=zotero&xid=7f71e179

12. Sanderson M, Bulloch AG, Wang J, Williamson T, Patten SB. 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. doi:10.1016/j.jad.2019.12.024

13. Jeblee S, Gomes M, Jha P, Rudzicz F, Hirst G. Automatically determining cause of death from verbal autopsy narratives. *BMC Med Inform Decis Mak*. 2019;19(1):127. doi:10.1186/s12911-019-0841-9

14. Zhang H. The Optimality of Naive Bayes. *FLAIRS Conference*. Published online 2004.

15. 1.9. Naive Bayes — scikit-learn 0.22.2 documentation. Accessed April 19, 2020. https://scikit-learn.org/stable/modules/naive_bayes.html

16. Liaw A, Wiener M. Classification and Regression by randomForest. 2002;2:5.

17. Nembrini S, König IR, Wright MN. The revival of the Gini importance? *Bioinformatics*. 2018;34(21):3711-3718. doi:10.1093/bioinformatics/bty373

18. Elements of Statistical Learning: data mining, inference, and prediction. 2nd Edition. Accessed February 12, 2020. https://web.stanford.edu/~hastie/ElemStatLearn/

19. Bradley B. UC Business Analytics R Programming Guide: Gradient Boosting Machines. Accessed June 4, 2020. http://uc-r.github.io/gbm_regression

20. Sze V, Chen Y-H, Yang T-J, Emer J. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *arXiv:170309039 [cs]*. Published online August 13, 2017. Accessed February 12, 2020. http://arxiv.org/abs/1703.09039

21. Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies | Population Health Metrics | Full Text. Accessed May 30, 2020. https://pophealthmetrics.biomedcentral.com/articles/10.1186/1478-7954-9-28

22. Murray C, Lopez A, Black R, et al. Population Health Metrics Research Consortium gold standard verbal autopsy validation study: Design, implementation, and development of analysis datasets. *Population health metrics*. 2011;9:27. doi:10.1186/1478-7954-9-27

23. Hooker S, Erhan D, Kindermans P-J, Kim B. A Benchmark for Interpretability Methods in Deep Neural Networks. In: Wallach H, Larochelle H, Beygelzimer A, Alché-Buc F d\textquotesingle, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc.; 2019:9737–9748. Accessed June 3, 2020. http://papers.nips.cc/paper/9167-a-benchmark-for-interpretability-methods-in-deep-neural-networks.pdf

24. Sanders H, Saxe J. Garbage In, Garbage Out: How Purportedly Great ML Models Can Be Screwed Up By Bad Data. *Proceedings of Blackhat 2017*. Published online July 17, 2017.