RESEARCH

Automatically determining cause of death from verbal autopsy narratives

Serena Jeblee[1*], Mireille Gomes[2], Prabhat Jha[2,3], Frank Rudzicz[4,1] and Graeme Hirst[1]

[*]Correspondence:
sjeblee@cs.toronto.edu
[1]Department of Computer Science, University of Toronto, Toronto, Canada
Full list of author information is available at the end of the article

Abstract

Background: A verbal autopsy (VA) is a post-hoc written interview report of the symptoms preceding a person's death in cases where no official cause of death (CoD) was determined by a physician. Current leading automated VA coding methods primarily use structured data from VAs to assign a CoD category. We present a method to automatically determine CoD categories from VA free-text narratives alone.

Methods: After preprocessing and spelling correction, our method extracts word frequency counts from the narratives and uses them as input to four different machine learning classifiers: naïve Bayes, random forest, a support vector machine, and a neural network.

Results: For individual CoD classification, our best classifier achieves a sensitivity of .770 for adult deaths, as compared to the current best reported sensitivity of .57. When predicting the CoD distribution at the population level, our best classifier achieves .962 cause-specific mortality fraction accuracy, which is on par with leading CoD distribution estimation methods.

Conclusions: Our narrative-based classifier substantially outperforms current machine learning classifiers at the individual level. Moreover, our method demonstrates that VA narratives provide important information that can be used by a machine learning system to achieve better CoD classification accuracy than the structured data alone.

Keywords: cause of death; computer-coded verbal autopsy (CCVA); physician certified verbal autopsy (PCVA); machine learning; natural language processing; Tariff method; verbal autopsy

1 Background

1.1 Verbal Autopsies

Two-thirds of the world's 60 million deaths each year do not have a known cause of death (CoD). The largest gap between known and unknown CoDs is in developing countries, where many deaths occur at home rather than in health facilities [1]. Verbal autopsy (VA) surveys can help to bridge this gap by providing information about the most prevalent causes, which helps to inform public health planning and resource allocation [2]. A VA survey typically involves interviews with family members of the deceased, conducted by non-medical staff who complete a structured questionnaire of symptoms and risk factors before death. They also ask the family members about the events and circumstances around the time of death and record the responses in a free-text narrative. Typically, two or more physicians review each completed VA survey and independently make a CoD diagnosis [3], with reconciliation done by another more senior physician if necessary.

Although there have been criticisms of physician-coded VAs [4], there is no gold standard for VA coding that we can evaluate against, since for most VAs we have no way of knowing the true CoD. Records of hospital deaths cannot be considered a gold standard for non-hospital deaths because of the differences in the distribution and characteristics of the patients who receive care in hospitals and those who die at home without medical attention (such as education level, access to hospital care, types of pathogens, etc.) [3, 5, 6]. For this reason, physician-coded VAs are often used for training and testing automated CoD coding methods.

Automated CoD coding may help to reduce physician time and costs when coding VA surveys. So far, machine learning techniques have been primarily applied to data from the structured questionnaires only, with the best sensitivity scores around .60 for individual CoD classification, using various numbers of CoD categories (typically 15–30) [7]. Some studies have suggested that the narrative section is unnecessary or of limited use for determining CoD [8]. However, we hypothesize that using the structured questions alone results

in insufficient accuracy because information that appears only in the free-text narrative is often essential to making a correct diagnosis, such as symptom chronology and treatment history [9]. Our method uses word frequency counts from the narrative to determine the appropriate CoD category for a VA record. We explore several different models including naïve Bayes, random forests, support vector machines, and a neural network.

1.2 Metrics

In the absence of medical death certification in low- and middle-income countries, VAs are primarily used to estimate the proportion of deaths from various causes at the population level, so as to inform public health planning. Subsequently, individual level VA CoD assignments are often aggregated to determine the CoD distribution in the population.

To evaluate CoD classification at the individual level, we report precision (positive predictive value), sensitivity (recall), and F_1-measure (the harmonic mean of precision and sensitivity), as well as partial chance-corrected concordance (PCCC). Chance-corrected correspondence (CCC) is a measure of how well the predicted CoD categories correspond to the correct CoD categories, and PCCC is the same measure adjusted for the number of possible categories [10]. To evaluate the CoD distribution prediction at the population level, we report Cause-Specific Mortality Fraction (CSMF) accuracy [10, 11]. CSMFs measure the relative proportions of CoDs in a population, and CSMF accuracy measures the similarity of the distribution of CoD categories assigned by the classifier to the true distribution.

However, CSMF accuracy scores of .50 or above can often be achieved by random guessing, especially if the method takes into account the training distribution. So we also report chance-corrected CSMF accuracy (CCCSMFA) [12], which produces a score of 0 for chance performance, and a negative score for performance worse than chance.

1.3 Previous work

Several expert-driven and machine learning methods have been used for automatically categorizing VAs by CoD, at both the individual and the population level [13, 14, 15, 16,

17, 18, 19]. Many of these methods are based on questionnaires such as the World Health Organization (WHO) 2016 Verbal Autopsy Instrument [20], which is a standardized VA questionnaire with detailed questions about the subject's symptoms and medical history.

Boulle et al. [13] were among the first to use neural networks for VA CoD classification in 2001. They used a small set of structured questionnaire data with a neural network and achieved a sensitivity of .453 for individual classification into 16 CoD categories. However, to our knowledge, no current VA coding method uses neural networks despite their recent popularity.

The King-Lu method [21] uses the conditional probability distributions of symptoms to estimate the CoD distribution of a dataset. It does not provide a CoD for individual records. Desai et al. [7] reported a CSMF accuracy of .96 using the King-Lu method on the Indian Million Death Study dataset [3].

InterVA-4, a popular automated VA coding method developed by Byass et al. [14], uses a predetermined list of symptoms and risk factors extracted from a structured questionnaire. Records are assigned a CoD based on conditional probabilities for each symptom given a CoD, as assigned by medical experts, as well as the probabilities of the CoDs themselves. Miasnikof et al. [17] reported a sensitivity of .43 and CSMF accuracy of .71 for InterVA-4 on data from the Million Death Study [3].

InSilicoVA, described by McCormick et al. [15], is a statistical tool that uses a hierarchical Bayesian framework to estimate the CoD for individual records as well as the population distribution. They reported a mean sensitivity of .341 across 34 CoD categories for individual records, and .85 CSMF accuracy.

The Tariff Method, presented by James et al. [16, 22], uses a sum of weighted scores (tariffs) to determine the most probable CoD. The score for each of the possible CoDs is the weighted sum of different tariffs, which are each calculated from the value of a certain indicator (usually a symptom or risk factor). Most of these indicators are taken from the structured questionnaire, although there are also tariffs that represent the presence of some

frequent narrative words (50 or more occurrences in the training data). James et al. reported .505 CCC and .770 CSMF accuracy for adult records from the Population Health Metrics Research Consortium (PHMRC) dataset[23], using 53 CoD categories.

Miasnikof et al. [17] used a naïve Bayes classifier to assign CoD categories. They evaluated their classifier on several different datasets, including the PHMRC dataset and the Million Death Study dataset [3, 24], which we will use in this paper (see section 3), with 16 CoD categories. They obtained results that surpassed those of the Tariff Method and InterVA-4, including a sensitivity of .57 and CSMF accuracy of .88. However, their model used only data from the structured questionnaire.

Danso et al. [18] used word frequency counts and tf·idf scores (the frequency of a term divided by the frequency of documents in which it occurs) from VA narratives as features (measurable characteristics of data that are used as input to computational models) with a support vector machine (SVM) classifier, achieving a maximum F_1 score of .419. They also used a naïve Bayes classifier and a random forest classifier, which achieved F_1 scores of .373 and .149 respectively. They did not report population level metrics.

Danso et al. [19] used a variety of linguistic features such as part-of-speech tags, noun phrases, and word pairs from 6,407 VA narratives of infant deaths from Ghana, and classified the records into 16 CoD categories, achieving a sensitivity of .406 using only the narrative-based features and .616 using a combination of narrative and structured questionnaire features. They noted that they achieved better performance with the linguistic features than with only word occurrence features, though their dataset was small and the part-of-speech tagger was not trained on medical data, and thus is likely to produce incorrect part-of-speech information.

2 Methods

2.1 Data

Our main dataset comes from the Million Death Study (MDS), the goal of which is to provide a national estimate of the leading CoDs in India in order to enable evidence-based

health programming [3, 24]. Since the majority of available records in MDS are scans of handwritten forms, which not usable by our automated prediction tool, we use a subset consisting of the records with narratives that have been transcribed into a digital format. It consists mostly of English narratives, which tend to come from southern and northeastern India. However, all states are represented in this dataset. The remaining narratives have been translated into English from various local languages. In addition to this dataset, we also have a set of records from a recent multi-centre randomized control trial (RCT) that was conducted in four districts within two states of India: Gujarat and Punjab, on 9,374 deaths [25]. The aim of this RCT was to assess whether current leading machine learning algorithms perform as well as physician diagnosis when determining the CoD for VAs at the population level. The RCT collected VAs on all deaths from the study sites up to age 70 that occurred within five years preceding the study. Approximately half of these deaths were randomly assigned for coding by physicians, for which VA structured questions and narratives were collected, and the remainder of the deaths were assigned to automated methods for coding using VA questionnaires with structured questions only. A randomly selected subset of the narratives from this RCT were translated into English, and are used in this study.

In the MDS and RCT datasets, each record is assigned a WHO International Classification of Diseases (ICD) version 10 code [26] by two specially trained physicians who independently and anonymously review each record. When the two assigned codes do not match (about 30% of records), the records undergo anonymous reconciliation, and persisting disagreements (about 15%) are adjudicated by a third senior physician. This process is standard for physician-coded VAs [20] and was conducted independently of developing our automated method.

In the combined datasets there are over 500 ICD-10 codes, so the records are grouped into 15 CoD categories for records of adult (15–69 years) and child (29 days–14 years) deaths, and 5 categories for records of neonatal (<29 days) deaths. See Tables 1 and 2 for

CoD categories, and additional file 1 for the complete mapping of ICD-10 codes to CoD categories.

In addition, we also train and test our models on the Agincourt dataset, which is composed of coded VA records of community deaths in South Africa [27]. See Table 3 for details of the datasets.

Since VA narratives are often handwritten and then transcribed and perhaps translated, there are frequent spelling errors and grammatical inconsistencies due to varying levels of experience of the surveyors and quality of the translations. In addition, medical symptoms are often described in non-technical or local terms by the non-medical surveyors. Due to the informal nature of the text and the frequency of errors, we focus on individual words, which avoids some of the issues with the grammatical inconsistencies. See Table 4 for some examples of narrative text from the MDS dataset.

2.2 Implementation of metrics

In order to evaluate chance-corrected CSMF accuracy, we conducted the Monte Carlo calculation described by Flaxman et al. [12] with 10,000 iterations, and found the mean CSMF accuracy of randomly assigning CoD categories to be .646 for the neonatal dataset (5 CoD categories), .641 for child dataset (15 CoD categories), and .643 for the adult dataset (15 CoD categories). We use these values as the mean for chance-correcting CSMF accuracy because they are specific to our dataset, although they are very close to the value of .632 that Flaxman et al. reported.

Since the records for each test set and training set are selected randomly, we expect the test distributions to be similar to the training distributions. Some VA studies have resampled their training and test set to create uniform distributions in order to avoid the model learning to assign CoD categories to individual records based on the frequency of the CoD categories [17, 22]. However, we chose not to do so because because some CoD categories have a very small number of records and achieving a reasonably sized test set

would require us to replicate some records many times, which would not constitute a fair evaluation of our method.

2.3 Machine learning models for text classification

Like the MDS, the RCT data was also collected in India and follows a similar protocol to the MDS [3, 24, 25], so the two sets were combined to create a bigger dataset to train and test our method with. Unlike these datasets, the Agincourt data was captured in South Africa and has greater variations in protocol [27], and hence was not combined with the other datasets. Early experiments showed that the model performed better with more training data, which is typical of machine learning classifiers. The datasets were preprocessed as follows. Spelling was corrected by using the PyEnchant Python library [28] with an English dictionary and a short hand-crafted dictionary containing common terms that appear in the narratives. The text was subsequently lowercased and punctuation separated from words. A set of 160 stopwords (such as *and*, *because*, *for*) were removed from the narratives.[1] The remaining words were stemmed (i.e. morphological endings removed) with the Porter Stemmer[2]; for example, the stem of *crying* is *cry*.

The features that we use for CoD classification are word frequency counts from the narrative and one feature that indicates whether the record is of an adult, child, or neonatal death. We compute the analysis of variance (ANOVA) F-value[3] for each feature, which calculates the ratio of the variance between the means of the feature values for each of the CoD categories, to the variance within each class. If the means are significantly different between CoD categories and the variance within categories is small, then the feature is likely to be discriminative. We keep only the highest scoring features, reducing the space from over 4000 to several hundred features, depending on the model.

For our classifiers, all models except the neural network are created in Python with scikit-learn [29]. Each classifier is optimized with hyperopt [30] for 100 runs for model parameters and the number of features, using a small subset of the MDS data. The models are optimized separately so that we are comparing the best version of each model. The naïve

Bayes classifier, which assigns a CoD category to a record using the independent conditional probabilities for each feature, uses the best 200 features. The random forest model, which uses a combination of learned decision trees to classify new data points, uses the best 414 features and 26 trees.

Support vector machines (SVMs) are commonly used models that learn to classify data by maximizing the margin between categories in the training data, using a kernel function that maps the input features to higher dimensional space. Our SVM model is an aggregate of one-vs-rest SVMs with linear kernel functions, using 378 features.

Neural networks are made up of layers of simulated neurons with connections between the layers that can transmit information. The neural network model we use is a feed-forward network with one hidden layer (297 nodes, chosen by optimization) created with Keras [31], using Theano [32] as the backend. It uses 398 features and rectified linear units (ReLUs) as the activation function (the function that computes the output of an artificial neuron in the network given input values and learned weights).

For the adult and child datasets, each training set is augmented with all the data from the other two datasets. In general, we found that the classifiers perform better with extra training data, especially for the smaller child dataset. For neonatal records, the models are trained only with neonatal data because these records use a different set of CoD categories.

3 Results

Table 5 shows the mean scores for each classifier using 10-fold cross-validation with the combined MDS and RCT data. Each of the 10 test splits contained approximately 1,204 adult records, 185 child records, and 57 neonatal records. Overall, the neural network performs the best in terms of sensitivity, with .770 for adults, .695 for child records, and .576 for neonatal records. However, for CSMF accuracy the best performance is achieved by the SVM and neural network classifiers on adult records (.962), and the SVM on child records (.914) and neonatal records (.857).

In comparison to our model's sensitivity of .770 for adult deaths and .695 for child deaths, Miasnikof et al. [17] reported a mean sensitivity of .57 on MDS checklist data from child and adult deaths with their naïve Bayes classifier and 16 CoD categories. They compared their results to InterVA-4 on the Million Death Study data, which achieved .43, and the Tariff Method, which achieved .50 sensitivity. InSilicoVA reported a sensitivity of .341 using 34 CoD categories for adult deaths from the PHMRC dataset [23]. Danso et al. [19] reported a sensitivity of .406 with their SVM classifier using narrative features from a dataset of 6407 neonatal records and 16 CoD categories, and .616 using narrative and structured data features, while our model achieved .576 sensitivity for records of neonatal deaths using only the narrative.

Our neural network classifier's CMSF accuracy score of .962 for adult deaths and .914 for child deaths surpasses all other methods except the King-Lu method on MDS data (.96) [7], but the King-Lu method does not assign CoD categories to individual records. Miasnikof reported a CSMF accuracy of .88 for their model, .71 for InterVA-4, .57 for the Tariff Method, and InsilicoVA reported .85 CSMF accuracy.

See Table 6 for results on the Agincourt dataset. As with the MDS dataset, the neural network performs the best for adult records, with a sensitivity of .578 and PCCC of .547. For the Agincourt neonatal records, the naïve Bayes model performs the best (.526 sensitivity and .404 PCCC), likely because the dataset is so small. By comparison, Miasnikof et al. [17] reported an overall sensitivity of .48 and PCCC of .43 on the Agincourt dataset, and Desai et al. reported a PCC of .38 using the open source Tariff method and .39 using InterVA-4.

4 Discussion

Some have suggested that it might be better to replace the free-text portions with more detailed checklist items to avoid the overhead of manually collecting, transcribing, translating, and processing the narrative [23]. While structured data can be very useful, it is more time-consuming to collect, and currently does not capture information such as chronology

and health-seeking behaviors that is often made available via the narrative. We have demonstrated that despite the varying quality of the narrative text, it can still be used to achieve high agreement with physician-determined CoD.

While most other methods achieve their results by using expert-driven features or a large amount of data from the structured questionnaire in addition to some narrative-based features (in the case of the Tariff method[22] and Danso et al.[18, 19]), our model uses only the narrative and thus can be trained and tested on any set of verbal autopsies that contain free-text narratives, and we are able to achieve better or comparable performance to previously reported automated methods.

A possible explanation for why our narrative-based classifiers performed better than that of Danso et al., especially the random forest model, is that not only did we train on more data, but we also performed feature selection and parameter optimization for each classifier, while Danso et al. only performed feature reduction for the SVM, and used the default parameters for all models. Better feature selection helps to prevent overfitting to the training data and reduce computation time for our models. Some of the highest ranked features that were selected by the ANOVA module are words like *yellow*, *abdomen*, *weak*, *fever*, *cough*, etc, which clearly describe symptoms. Some of the features seemed to describe conditions or situations, such as *pregnancy*, *cancer*, and *tuberculosis*, and some were less obvious, such as *help*, *gradually*, and *one*.

Certain CoD categories have fewer misclassifications, most notably “Suicide” and “Road and transport injuries”. Those narratives tend to be less complex since the CoD is well identified within the text. The most commonly confounded CoD categories were “Other non-communicable diseases” and “Ill-defined”. The classifiers seem to have more trouble distinguishing between CoD categories that have a large variation in symptom patterns, which are also more difficult for humans to diagnose.

One disadvantage of our method is that some narratives are long and include background information that is not ultimately relevant to the CoD, such as a history of smoking or

asthma when the subject died in a car accident. Sometimes the respondents mention what they believe to be the CoD in the narrative, which may or may not be the CoD that is subsequently determined by the physicians. The presence of those words in the narrative could potentially cause a misclassification. In addition, the word frequency counts do not take into account word order, and consequently, higher-level linguistic information such as negation and chronology is not captured. We plan to handle some of these issues in the future by using models that capture the sequence of the words, and we also plan to use temporal relation extraction to account for chronology. However, the present work provides a strong baseline for narrative-based automated VA coding.

5 Conclusions

We have shown that a variety of narrative-based machine learning classifiers can be used for automated VA coding. Unlike most other methods, ours does not rely on a specific structured data format or questionnaire; it can be applied to any English VA narrative, and is more adaptable to different datasets and populations than methods that rely on structured data.

No current method for automatically determining CoD for VA records has sufficient accuracy to be a replacement for human doctors. However, we have shown that for adult deaths, the largest group of deaths in our dataset, that our method can achieve .770 sensitivity and over .90 agreement (CSMF accuracy) at the population level with physician-assigned CoDs, which is significantly better than any other current method. This demonstrates that narrative-based machine learning methods are a promising option for CoD coding for VAs. A large repository of openly available VA data with full narratives and physician-assigned cause of death would help in further development of such computational methods.

To improve this method, we are currently considering combinations of features from the structured data and the narrative in order to produce an automated CoD coding tool that is robust and reliable enough to be used in the field. In our ongoing work, we are using more

linguistically motivated features that take into account context, chronology, and semantics, and we are also exploring alternative neural network architectures.

Abbreviations

ANOVA: analysis of variance, CCC: chance-corrected correspondence, CCCSMFA: chance-corrected CSMF accuracy, CSMF: cause-specific mortality fraction, CoD: cause of death, ICD: International Classification of Diseases, MDS: Million Death Study, PCCC: partial chance-corrected concordance, PHMRC: Population Health Metrics Research Consortium, RCT: randomized control trial, SVM: support vector machine, VA: verbal autopsy, WHO: World Health Organization

Declarations

Ethics approval and consent to participate

Ethics approval for the Million Death Study was obtained from the Post Graduate Institute of Medical Research, St. Johns Research Institute and St. Michaels Hospital, Toronto, Ontario, Canada.

Ethical clearance for health and demographic surveillance in Agincourt was granted by the University of the Witwatersrands Committee for Research on Human Subjects (Medical).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The MDS and RCT datasets are the property of the Government of India and cannot be shared.

Funding

This work was funded by a Google Faculty Research Award, the US National Institutes of Health, and the University of Toronto.

Author's contributions

SJ performed the data preprocessing and implemented the algorithms and evaluation. SJ wrote the paper with guidance from MG and GH. MG provided the data as well as guidance about the methods, evaluation, and background information. GH and FR provided guidance about the computational methodology and evaluation. PJ oversaw the MDS data collection. All authors contributed to data interpretation and critical revisions of the paper. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by grants from Google, the NIH and the University of Toronto.

Figures

Figure 1: **Adult CoD category distribution**

Figure 2: **Child CoD category distribution**

Figure 3: **Neonate CoD category distribution**

Tables

Table 1: CoD categories used for adult deaths (15–69 years), and child deaths (29 days–14 years)

Acute respiratory infections
Diarrhea
Pulmonary Tuberculosis
Other and unspecified infections
Neoplasms
Nutrition
Cardiovascular disease
Chronic respiratory disease
Liver cirrhosis
Other non-communicable diseases
Road and transport injuries
Other injuries
Ill-defined
Suicide
Maternal

Table 2: CoD categories used for neonatal deaths (<29 days)

Prematurity/low birth weight
Neonatal infections (not including tetanus)
Birth asphyxia/trauma
Ill-defined or cause unknown
Other (all other ICDs not included in above)

Table 3: Description of datasets used. MDS: Million Death Study dataset, RCT: Randomized Control Trial dataset

	MDS	**RCT**	**MDS+RCT**	**Agincourt**
Adult records (15–69 years)	9,215	2,830	12,045	8,151
Child records (29 days–14 years)	1,721	130	1,851	1,674
Neonatal records (<29 days)	465	107	572	197
Region	India	India (Gujarat, Punjab)	India	South Africa

Table 4: Two example narratives (adult deaths)

Narrative	**Physician certified CoD category**
Heart failure. The patient death due to breathlessness. The person suffering paralysis and stroke lost on year with chest pain very pressure after then person was head.	Cardiovascular disease
One day 13/03/01 he fell ill with some fever and chest pain who called the Doctor. On 15/03/01 the deceased was crying in the chest pain and high fever. We were ready to shift. The patient to the Hospital, some water came out from the deceased mouth and closed his eyes and passed away.	Acute respiratory infections

Table 5: Mean scores on the combined MDS and RCT datasets for each of the four classifiers. Adult and child results classified into 15 categories; neonatal records into 5 categories. Bold indicates the best score in each column for each age group. PCCC: partially chance-corrected concordance, CSMFA: cause-specific mortality fraction (CSMF) accuracy, CCCSMFA: chance-corrected CSMFA

Adult (15–69 years)	**Precision**	**Sensitivity**	**F_1**	**PCCC**	**CSMFA**	**CCCSMFA**
Naïve Bayes	.710	.710	.704	.689	.929	.801
Random forest	.733	.730	.728	.711	.948	.854
SVM	.746	.737	.740	.718	**.962**	**.894**
Neural network	**.773**	**.770**	**.770**	**.764**	**.962**	**.894**
Child (29 days–14 years)	**Precision**	**Sensitivity**	**F_1**	**PCCC**	**CSMFA**	**CCCSMFA**
Naïve Bayes	.647	.595	.608	.565	.851	.585
Random forest	.687	.620	.638	.591	.872	.643
SVM	.686	.658	.666	.632	**.914**	**.760**
Neural network	**.719**	**.695**	**.698**	**.672**	.904	.733
Neonate ($<$29 days)	**Precision**	**Sensitivity**	**F_1**	**PCCC**	**CSMFA**	**CCCSMFA**
Naïve Bayes	.507	.516	.493	.376	.826	.509
Random forest	.534	.542	.524	.411	.852	.581
SVM	.537	.538	.524	.404	**.857**	**.597**
Neural network	**.579**	**.576**	**.556**	**.453**	.825	.507

Table 6: Mean scores on the Agincourt dataset. CCCSMFA was calculated using .632 as the mean of random allocation, as suggested in [12].

Adult (15–69 years)	**Precision**	**Sensitivity**	**F_1**	**PCCC**	**CSMFA**	**CCCSMFA**
Naïve Bayes	.517	.517	.513	.481	**.932**	**.814**
Random forest	.511	.517	.496	.480	.844	.577
SVM	.569	.566	.561	.543	.901	.730
Neural network	**.575**	**.578**	**.570**	**.547**	.918	.777
Child (29 days–14 years)	**Precision**	**Sensitivity**	**F_1**	**PCCC**	**CSMFA**	**CCCSMFA**
Naïve Bayes	.488	.440	.435	.395	.761	.351
Random forest	.521	.502	.487	.463	.816	.501
SVM	.535	.518	.512	.479	**.872**	**.653**
Neural network	**.572**	**.562**	**.552**	**.527**	.869	.645
Neonate ($<$29 days)	**Precision**	**Sensitivity**	**F_1**	**PCCC**	**CSMFA**	**CCCSMFA**
Naïve Bayes	**.532**	**.526**	**.483**	**.404**	.702	.191
Random forest	.409	.496	.427	.366	**.710**	**.213**
SVM	.387	.417	.371	.266	.693	.165
Neural network	.356	.412	.354	.259	.636	.012

Additional files

Additional file 1: Cause of death categories with corresponding ICD-10 codes (pdf)

Notes

[1]Danso et al. [18] also lowercased the text in their dataset but removed punctuation and did not remove stopwords or perform spelling correction.

[2]We use the implementation of the Porter Stemmer provided in NLTK [33].

[3]We use scikit-learn's SelectKBest module with the f_classif function [29].

Author details

[1]Department of Computer Science, University of Toronto, Toronto, Canada. [2]Centre for Global Health Research, St. Michael's Hospital, Toronto, Canada. [3]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada. [4]Toronto Rehabilitation Institute-UHN, Toronto, Canada.

References

1. Department of Economic and Social Affairs, Population Division, United Nations. World Population Prospects: The 2012 revision. ST/ESA/SER.A/334; 2013.
2. Jha P. Reliable direct measurement of causes of death in low- and middle-income countries. BMC Medicine. 2014;12:19.
3. Aleksandrowicz L, Malhotra V, Dikshit R, Prakash C Gupta RK, Sheth J, Rathi SK, et al. Performance criteria for verbal autopsy-based systems to estimate national causes of death: Development and application to the Indian Million Death Study. BMC Medicine. 2014;12:21.
4. Lozano R, Lopez AD, Atkinson C, Naghavi M, Flaxman AD, Murray CJ. Performance of physician-certified verbal autopsies: multisite validation study using clinical diagnostic gold standards. Population Health Metrics. 2011;9(32).
5. Ram U, Dikshit R, Jha P. Level of evidence of verbal autopsy–Authors' reply. The Lancet Global Health. 2016;4(6):e368–e9.
6. Berkley JA, Lowe BS, Mwangi I, Williams T, Bauni E, Mwarumba S, et al. Bacteremia among children admitted to a rural hospital in Kenya. New England Journal of Medicine. 2005;352(1):39–47.
7. Desai N, Aleksandrowicz L, Miasnikof P, Lu Y, Leitao J, Byass P, et al. Performance of four computer-coded verbal autopsy methods for cause of death assignment compared with physician coding on 24,000 deaths in low- and middle-income countries. BMC Medicine. 2014;12:20.
8. King C, Zamawe C, Banda M, Bar-Zeev N, Beard J, Bird J, et al. The quality and diagnostic value of open narratives in verbal autopsy: A mixed-methods analysis of partnered interviews from Malawi. BMC Medical Research Methodology. 2016;16:13.
9. Gajalakshmi V, Peto R. Commentary: Verbal autopsy procedure for adult deaths. International Journal of Epidemiology. 2006;35(3):748–750.
10. Murray CJ, Lozano R, Flaxman AD, Vahdatpour A, Lopez AD. Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. Population Health Metrics. 2011;9:28. Erratum [11].
11. Murray CJ, Lozano R, Flaxman AD, Vahdatpour A, Lopez AD. Erratum To: Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. Population Health Metrics. 2014;12:7.
12. Flaxman AD, Serina PT, Hernandez B, Murray CJ, Riley I, Lopez AD. Measuring causes of death in populations: a new metric that corrects cause-specific mortality fractions for chance. Population Health Metrics. 2015;13:28.
13. Boulle A, Chandramohan D, Weller P. A case study of using artificial neural networks for classifying cause of death from verbal autopsy. International Journal of Epidemiology. 2001;30(3):515–520.
14. Byass P, Chandramohan D, Clark S, D'Ambruoso L, Fottrell E, Graham W, et al. Strengthening standardised interpretation of verbal autopsy data: The new InterVA-4 tool. Global Health Action. 2012;5:19281.
15. McCormick TH, Li ZR, Calvert C, Crampin AC, Kahn K, Clark S. Probabilistic cause-of-death assignment using verbal autopsies. Journal of the American Statistical Association. 2016;111(15):1036–1049.
16. James SL, Flaxman AD, Murray CJ. Performance of the Tariff Method: Validation of a simple additive algorithm for analysis of verbal autopsies. Population Health Metrics. 2011;9(1):31–47.
17. Miasnikof P, Giannakeas V, Gomes M, Aleksandrowicz L, Shestopaloff AY, Alam D, et al. Naïve Bayes classifiers for verbal autopsies: Comparison to physician-based classification for 21,000 child and adult deaths. BMC Medicine. 2015;13(1):286–294.
18. Danso S, Atwell E, Johnson O. A comparative study of machine learning methods for verbal autopsy text classification. International Journal of Computer Science Issues. 2013;10(6).

19. Danso S, Atwell E, Johnson O. Linguistic and Statistically Derived Features for Cause of Death Prediction from Verbal Autopsy Text. In: Language Processing and Knowledge in the Web. Springer Berlin Heidelberg; 2013. p. 47–60.
20. Nichols EK, Byass P, Chandramohan D, Clark SJ, Flaxman AD, Jakob R, et al. The WHO 2016 verbal autopsy instrument: An international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0. PLOS Medicine. 2018 01;15(1):1–9.
21. King G, Lu Y. Verbal autopsy methods with multiple causes of death. Statistical Science. 2008;23(1):78–91.
22. Serina P, Riley I, Stewart A, James SL, Flaxman AD, Lozano R, et al. Improving performance of the Tariff Method for assigning causes of death to verbal autopsies. BMC Medicine. 2015 Dec;13(1):291.
23. Population Health Metrics Research Consortium (PHMRC). Population Health Metrics Research Consortium Gold Standard Verbal Autopsy Data 2005-2011; 2013. http://ghdx.healthdata.org/record/population-health-metrics-research-consortium-gold-standard-verbal-autopsy-data-2005-2011.
24. Gomes M, Begum R, Sati P, Dikshit R, Gupta PC, Kumar R, et al. Nationwide Mortality Studies to Quantify Causes of Death: Relevant Lessons from India's Million Death Study. Health Affairs. 2017;36(11):1887–1895.
25. Gomes M, Kumar D, Budukh A, et al. Computer versus Physician Coding of Cause of Deaths using Verbal Autopsies: a randomised trial of 9374 deaths in four districts of India;In press.
26. World Health Organization. International statistical classifications of diseases and related health problems. 10th rev. vol. 1. Geneva, Switzerland: World Health Organization; 2008.
27. Kahn K, Collinson M, Gmez-Oliv F, Mokoena O, Twine R, Mee P, et al. Profile: Agincourt health and socio-demographic surveillance system. International Journal of Epidemiology. 2012;41(4):988–1001.
28. Kelly R. PyEnchant; 2015. http://pythonhosted.org/pyenchant/.
29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(Oct):2825–2830.
30. Bergstra J, Yamins D, Cox DD. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: Proceedings of the 30th International Conference on Machine Learning (ICML 2013); 2013. p. 115–123.
31. Chollet F. Keras. GitHub; 2015. https://github.com/ fchollet/keras.
32. Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints. 2016 May;abs/1605.02688. http://arxiv.org/abs/1605.02688.
33. Bird S, Klein E, Loper E. Natural Language Processing with Python. O'Reilly Media; 2009.