Artificial Intelligence In Medicine 143 (2023) 102622

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Research paper

Cause of Death estimation from Verbal Autopsies: Is the Open Response redundant or synergistic?

Ander Cejudo [a], Arantza Casillas [a,*], Alicia Pérez [a], Maite Oronoz [a], Daniel Cobos [b,c]

[a] HiTZ Basque Center for Language Technologies - Ixa NLP Group, University of the Basque Country (UPV/EHU), Spain[1]
[b] Swiss Tropical and Public Health Institute, Switzerland
[c] University of Basel, Basel, Switzerland

ARTICLE INFO

Keywords:
Verbal autopsy
Natural language processing
Transformers
Cause of death

ABSTRACT

Civil registration and vital statistics systems capture birth and death events to compile vital statistics and to provide legal rights to citizens. Vital statistics are a key factor in promoting public health policies and the health of the population. Medical certification of cause of death is the preferred source of cause of death information. However, two thirds of all deaths worldwide are not captured in routine mortality information systems and their cause of death is unknown. Verbal autopsy is an interim solution for estimating the cause of death distribution at the population level in the absence of medical certification. A Verbal Autopsy (VA) consists of an interview with the relative or the caregiver of the deceased. The VA includes both Closed Questions (CQs) with structured answer options, and an Open Response (OR) consisting of a free narrative of the events expressed in natural language and without any pre-determined structure. There are a number of automated systems to analyze the CQs to obtain cause specific mortality fractions with limited performance. We hypothesize that the incorporation of the text provided by the OR might convey relevant information to discern the CoD.

The experimental layout compares existing Computer Coding Verbal Autopsy methods such as Tariff 2.0 with other approaches well suited to the processing of structured inputs as is the case of the CQs. Next, alternative approaches based on language models are employed to analyze the OR. Finally, we propose a new method with a bi-modal input that combines the CQs and the OR. Empirical results corroborated that the CoD prediction capability of the Tariff 2.0 algorithm is outperformed by our method taking into account the valuable information conveyed by the OR. As an added value, with this work we made available the software to enable the reproducibility of the results attained with a version implemented in R to make the comparison with Tariff 2.0 evident.

1. Introduction

Mortality statistics are essential for countries to inform health policies and design interventions to tackle the peaks of disease in the population. The final aim of producing cause of death statistics enables the comparison of changing health situations between countries. Having reliable and timely mortality statistics is necessary for an effective design and implementation of preventive interventions and health services. This implies that all countries should count with a health system with the required resources, both human and material, to identify the Cause of Death (CoD) of a considerable part of the population. That is not the case for all countries as many of them lack access to medical certification in many places within the country.

In areas where physicians are not available to certify the CoD, Verbal Autopsy (VA) has been shown to reliably provide this information at the population level [1]. A VA consists of a series of questions about the signs, symptoms, demographic characteristics and the condition that led to death answered by the relatives or the caregiver of the deceased. The VA instrument includes Closed Questions (CQ), and an Open Response (OR) where the interviewees can talk freely about how the death occurred. Expert clinicians are able to discern the most probable CoD from the answers conveyed in the VA. Needless to say, manual inspection of VA is time-consuming for expert clinicians.

In order to extract the CoD, VAs can be analyzed by either physicians or automated algorithms. Given the scale of some VA implementations, manual coding of VAs is becoming an unrealistic burden for

* Correspondence to: Ixa taldea, UPV-EHU, Manuel Lardizabal Ibilbidea, 1, Donostia-San Sebastián 20018, Spain.
E-mail address: arantza.casillas@ehu.eus (A. Casillas).
[1] (www.ixa.eus).

https://doi.org/10.1016/j.artmed.2023.102622
Received 19 September 2022; Received in revised form 19 May 2023; Accepted 1 July 2023
Available online 9 July 2023

many countries and the use of automated methods is gaining traction. To delve into the quantitative assessment of the VA, the InterVA [2] method was developed. This was the first method developed for VA analysis [3] and it mimics the behavior and decisions of a physician in terms of a heuristic algorithm based on hand-crafted rules. Then, the Institute for Health Metrics and Evaluation (IHME), along with other organizations, conducted the Population Health Metrics Research Consortium (PHMRC) gold standard verbal autopsy validation study [4]. As a result of this study, an VA gold-standard was released and made publicly available. Since then, some additional tools and algorithms have been implemented to automatically estimate the CoD given the CQs [5].

One of the algorithms tested in the PHMRC data was Tariff 2.0 [6], indeed, one of the WHO standards. The core idea of this algorithm is to assign a score, that is, a 'tariff' score to an item in the questionnaire according to the number of 'yes' given by the respondent for a certain CoD. The InSilicoVA method [7] also analyzes the VA data by identifying the most likely joint probability distribution of cause-specific mortality fractions. All these methods have in common that they estimate the CoD by only taking into account the CQs, while the OR is disregarded or used minimally.

In this work we explore the potential of the OR to contribute to the ascertainment of the CoD and our first **goal** is to quantitatively assess the predictive capabilities of the OR in comparison to the CQs. We hypothesize that the OR brings valuable information that should not be disregarded by automated methods. In addition to this, we explore if CQs and OR are redundant or whether they provide valuable information for CoD ascertainment. If the hypothesis is true, we could simplify the complex hierarchical questionnaire. If both CQs and OR would complement each other, this would imply that we could gain accuracy in the prediction of the CoD.

2. Previous work

The PHMRC data-set is not the only known VA data, for example, the Million Death Study (MDS) [8] collected the VA of thousands of deaths that occurred in India. There is also a data-set collected in Ghana [9] and another in Malasya [10], among others. The lack of available data could be due to the fact that VAs might convey sensitive information. PHMRC is, as far as we know, the only one publicly available.

There are **standards** which are recommended and gathered in the WHO 2016 instrument with the aim of seizing and assigning the most likely CoD to each VA. Li et al. [11], used the PHMRC data and the OpenVA [12] toolkit, which includes algorithms such as InterVA and InSilicoVA, and have reported a 21.24% accuracy for InterVA with the highest performance being 37.77% for the Naive Bayes [13] algorithm. Flaxman et al. [14] carried out a very similar study, but only with InSilicoVA, reporting a maximum accuracy of 37.6% for CoD assignment. McCormick et al. [5] compared also the algorithms inside the WHO 2016 instrument, using as input the PHMRC data-set and top 3 CoD accuracy evaluation.

Complementary studies have also analyzed verbal autopsy data with methods from the WHO 2016 instrument mainly by means of artificial intelligence techniques. Typically, these studies have as input the CQs, and less frequently, the OR. For OR, Danso et al. [15], by focusing on a data-set of VAs collected in Ghana, employed simple text representation methods with classical classification approaches. The Support Vector Machine (SVM) [16] classifier attained the highest macro-averaged F1 score [17], 41.9% with the TF-IDF [18] representation, to be precise. In contrast, Yan et al. [19], used word and character embeddings with the MDS data-set and they achieved 75.1% in F1 score metric for the adult age group. For CQs, Moran et al. [20] split and tested different subsets of the PHMRC data-set and outperformed the results of some of the WHO 2016 instrument algorithms, such as InSilicoVA, with their own Bayesian classifier (named BF). Li et al. [21], in similar comparison opted for a Gaussian mixture [22] that seemed to convey further improvements to the WHO 2016 algorithms.

The aforementioned studies made use of the information provided by the CQs, while the OR remains nearly unexplored. In addition, approaches that combine all the information are barely found. A combination could provide insights which would enable a simplification of the VA interview and help to improve the analysis.

3. Materials

This work makes use of the Verbal Autopsy Golden Standard (VA-GS) generated as a result of the PHMRC study [4]. In order to learn from VAs and evaluate the performance of the generated models, the data is split into train and test sets with stratified subsets with 70% and 30% of the VAs, respectively. In Table 1 a description of the whole data-set is shown. Together there are above 7400 samples randomly divided into train and test subsets with stratification. Each VA is described with two types of features (OR and CQs) and has a unique CoD assigned out of a total of 48 possible CoDs. The CQs are mainly categorical. Regarding the OR, the average length is 90, 75 and 87 for adult, child and neonate categories, respectively. Nonetheless, ORs with less than 10 words and over 200 words can also be found in the data-set. The dataset contains as the label the so called underlying cause of death which represents the condition that directly triggered the chain of events that led to death. Naturally, some CoDs are more frequent than others. For instance, 524 VAs were cases of pneumonia, while only 20 VAs had epilepsy as the CoD. On average, there are 109 VAs per CoD and, per age group, 100 for adults, 45 for children and 146 for neonate. Nevertheless, the deviation from the average is high as the CoD class is unbalanced. That is to say, there are different CoDs and the number of VAs per CoD is significantly different. For that reason, it is important to keep the same proportion of the CoD in both train and test sets. To this end, the train and test subsets were randomly selected with stratification. The split is the same as in [23] in an attempt to enable comparisons in Table 4.

Note that the CQs were specially designed for each age-segment (adult, child and neonate) since some questions do not apply to all the segments. Hence, the number of closed questions varies by segment; this is why in Table 1, there are 142 for adults (140 categorical and 2 numerical), 86 for children and 109 for neonates.

A large number of CQs lack value due to two phenomena: the presence of values such as "Don't know" and the so-called skip-patterns (e.g. questions that were not asked as they are considered unnecessary given the previous answers or contexts). These skip patterns are applied in some cases when, for a determinate CQ, a "Don't know" is answered and some of the subsequent CQs that follow up in the questionnaire are not asked, thus also obtaining a "Don't know" value. Another case is when, for example, the deceased is a man and the questions are designed to be answered by a woman. For instance, one of the CQs of the questionnaire is "Was the deceased a singleton or a multiple birth?" and if the answer is "Dont' know", it makes sense to put the same value in the next question of the questionnaire: "Was this the first, second or later in the birth order?". In addition, the CQs have on average 3 to 4 possible answers.

In Table 2, two examples of ORs and the corresponding CoDs are shown. As can be seen, the type of language used is non-technical and the corresponding CoD can be easily extracted from the first example. The CoD corresponds to an ICD-10 (International Classification of Diseases) code from a finite list of possible codes for this task, and which has a descriptor associated with it, for example, "Pneumonia". In the second example, the first two words do not add any useful information and they should not appear as there is more content in the open response. This happens throughout some of the ORs, where many texts have errors and others do not give any information, while other ORs mention the CoD explicitly.

Table 1
Quantitative description of the VA-GS data-set. Each verbal autopsy is an instance described in terms of bi-modal input information: i.e. an Open Response (OR) or free text and a Closed Questions (CQs); each verbal autopsy has annotated the Cause of Death (CoD), that is, the desired output. Out Of Vocabulary (OOV) tokens count the number of words of the vocabulary that appear in the test subset but not in the train subset.

VA-GS data-set				Adult	Child	Neonate	Total
Train	**Sample instances**			3,389	945	875	5,209
	Input: Features	OR	|Vocabulary|	8,056	3,059	3,284	9,540
			# Words	306,189	71,759	76,293	454,238
		CQ	Categorical	140	84	108	303
			Numerical	2	2	1	3
	Output: Class	CoD	Count	34	21	6	48
Test	**Sample instances**			1,460	396	377	2,233
	Input: Features	OR	|Vocabulary|	5,407	2,358	2,189	6,379
			# Words	131,068	29,264	29,893	190,225
			OOV	1,384	612	549	1,641
		CQ	Categorical	140	84	108	303
			Numerical	2	2	1	3
	Output: Class	CoD	Count	34	20	6	48

Table 2
Examples of two ORs with the corresponding CoD.

Open response	CoD
Father set on fire both mother and baby. Baby died in the afternoon. Father was in love with some other lady and to get married to her he killed his wife and baby.	Fires
No comments. They were twins. The boy died because his lungs were not developed, and had respiratory problems.	Preterm delivery

4. Methods

This section presents the methodological approach employed to estimate the CoD. As shown in Table 1 each VA is characterized by means of a bi-modal input $(CQ, OR) = (\mathbf{x_1}, \mathbf{x_2})$ and the aim is to estimate a CoD as the output of the approach. In this section we present different approaches to exploit the dual input. Each modality of the input shall be characterized by a feature-vector, respectively, of size m and n as in (1).

$$\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}) \in \mathbb{F}^m \times \mathbb{R}^n \text{ with} \tag{1}$$

$$\mathbf{x_1} = (x_{11}, \ldots, x_{1m}) \in \mathbb{F}^m \text{ to represent the CQs} \tag{2}$$

$$\mathbf{x_2} = (x_{21}, \ldots, x_{2n}) \in \mathbb{R}^n \text{ to represent the OR} \tag{3}$$

In Section 4.1, the inference of the CoD given the CQs is presented, while in 4.2, the results by the OR are shown. Finally, in 4.3, the **proposed** approach to deal with the dual OR and CQs is presented.

4.1. Models based on closed questions

In this section we explore approaches well suited for the CQs i.e. able to infer classifiers from either categorical or numerical feature-vectors (denoted, for simplicity, as in an m-sized space $\mathbb{F}^m$). In general, we shall refer to these methods as f_{CQ2CoD}, as in (4), due to the fact that they are able to compute the likelihood of the ith CoD (y_{1i}) given, as input, the responses to the CQs ($\mathbf{x_1} \in \mathbb{F}^m$). Thus, each y_{1i} is bound to the interval $[0, 1]$.

$$f_{CQ2CoD} : \mathbb{F}^m \longrightarrow [0,1]^{|CoD|} \qquad \mathbf{x_1} \longrightarrow f_{CQ2CoD}(\mathbf{x_1}) = (y_{11}, \ldots, y_{1|CoD|}) = \mathbf{y_1} \tag{4}$$

Typically, the estimated CoD, formally $\hat{c}_{CQ2CoD}$, is the most likely CoD, as in (5).

$$\hat{c}_{CQ2CoD} = \arg \max_{i=1}^{|CoD|} y_{1i} \tag{5}$$

Among the CQ2CoD approaches we also have, as well, with the **standard** Tariff 2.0 included in the WHO 2016 instrument [24]. The Tariff 2.0 algorithm is available in the OpenVA package [12]. This package is implemented in R and it offers functionalities such as downloading the VA data, parsing the information into different formats and also functions to train and assess the performance for CoD estimation in VAs. Tariff 2.0 has been designed to assign a score, that is, a 'tariff' score to an item in the questionnaire according to the count of 'yes' given by the respondent for a certain CoD. Tariff 2.0 identifies the strength of association between a symptom and a specific CoD, assuming that a symptom is statistically associated with certain diseases (e.g. a cough is associated with major respiratory diseases rather than others such as heart attack). This method outputs scores that are positive but not bounded to any interval, thus, a softmax layer is included in order to obtain probabilities for each CoD [25].

In addition, beyond score-based methods, we could rely on data-driven supervised inference approaches. Among them, we consider XGBoost [26], a gradient boosting algorithm. This algorithm is based on an ensemble of decision trees in a sequence, where each decision tree is adapted to minimize the errors made by the previous tree as it has as the input the output provided by the previous tree. The sequential adding of decision trees is done until the difference between the predicted CoD and the expected CoD reaches a minimum, in the so-called gradient descent. A variety of parameters can be set in order to maximize the performance for the specific task that it is used for: number of iterations (i.e. the number of trees to use), the maximum depth of the decision trees that are inferred during the training process and η as the learning rate.

Both methods included in the CQ2CoD approach provide, as the output, the likelihood (i.e. a number bound to $[0, 1]$) of every CoD to be the expected class of the input VA.

In addition, it should be taken into account that the questionnaire differs per age group. This implies that the value of m, the size of the input in (4) varies by age-segment and, accordingly, the function f_{CQ2CoD} is adapted to each age-segment.

4.2. Models based on open responses

As a second approach, we explored the OR as a source of valuable information to ascertain the CoD as in (6).

$$f_{OR2CoD} : \mathbb{R}^n \longrightarrow \mathbb{R}^{|CoD|} \qquad \mathbf{x_2} \longrightarrow f_{OR2CoD}(\mathbf{x_2}) = (y_{21}, \ldots, y_{2|CoD|}) = \mathbf{y_2} \tag{6}$$

Again, the estimated CoD, $\hat{c}_{OR2CoD}$, is the most reliable CoD, as in (7).

$$\hat{c}_{OR2CoD} = \arg \max_{i=1}^{|CoD|} y_{2i} \tag{7}$$

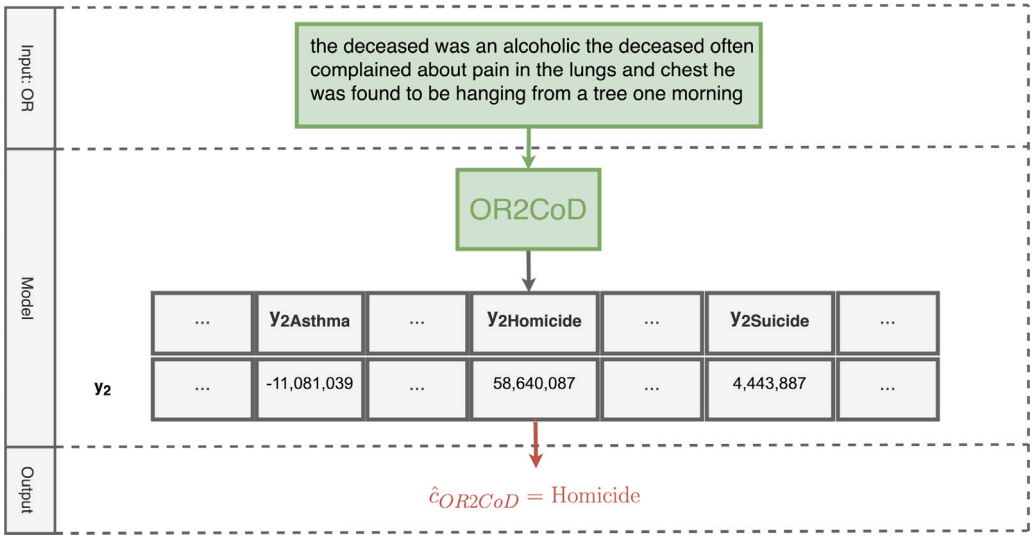

Fig. 1. Example of an incorrect estimation. On the top of the figure we have the Input (OR), and the expected or actual CoD is Suicide. For that input in the test set BERT predicted the output $\widehat{CoD}$ = Homicide as it was the CoD with the highest output value among the all possibles outcomes.

Fig. 1 shows an example of the output of this approach. The input to the model is the OR (a short text) and the output is an array ($\mathbf{y_2}$) with weights related to the reliability of each code, as denoted in (6). For the input string given in the example, the system estimates that the CoD 'Asthma' shows a reliability of $y_{2Asthma} = -11.0$ and, thus, is less reliable than 'Suicide' (with $y_{2Suicide} = 4.4$) but 'Homicide' has the highest reliability (with $y_{2Homicide} = 58.6$). Then, as in (7) the model provides the most reliable code ($\hat{c}_{OR2CoD}$ is 'Homicide' in the example). Note that, in the example, the expected or gold CoD was, by contrast, 'Suicide', meaning that the system failed in its estimation.

A key issue of this approach is the conversion of the free narrative into a meaningful numeric feature-vector ($\mathbf{x_2}$). Classical techniques like Bag of Words (BoW) [27] and TF-IDF [18] make such a vectorization possible. Nevertheless, these simple approaches do not convey contextual information as the strings are represented by a mere count of the presence of words in the narrative. Besides, these representations suffer from a high dimension (large n) due to the fact that the vocabulary involved in free narratives tends to be large (see the vocabulary involved in our task in Table 1). Instead, word embedding [28] is one of the emerging successful methods being currently employed to represent the text with numeric vectors. The added value of these methods rests on the fact that they encompass the contextual information of symbols (words) into dense numeric vectors of small dimension (small n).

Admittedly, having represented the OR in a vector, the information provided by the OR can be explored with classical data mining approaches such as Support Vector Machines [16], Logistic Regression [29], XGBoost itself and Recurrent Neural Networks (RNNs) [28] among others. Beyond the classical approaches, recent models like RNNs [28] and transformers [30] are suited to learn the features i.e. the word embeddings to effectively represent the input text. Moreover, the transformers possess attention mechanisms that learn to focus on different parts of the input data in order to accurately interpret the contextual information. Current trends in natural language understanding rest on language modeling as a means of keeping contextual information rather than mere symbolic representations. Transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) [31] are gaining importance for solving difficult tasks that require looking at the contextual information implicit in the language. BERT is composed of stacked encoder layers which have, mainly, two functions: several heads and a feed forward layer. An attention mechanism is called a head and it is the one that focuses more on some words rather than orders to maximize the final prediction accuracy. The information conveyed in the language model can be extrapolated in down-stream tasks involving language understanding. In this case, pre-trained transformers seem appropriate to cope with language modeling,while by fine-tuning the system learns to estimate the CoD.

In this work a variety of transformer-based models were considered and included among the OR2CoD approaches: Small BERT, BERT [31] and BioClinical BERT [32]. These transformer-based models do not provide a probability; that is, y_{2i} elements in (6) are not bound to $[0, 1]$. Instead, they give weights for each CoD that could be negative, as in the example provided in Fig. 1 Small BERT and BERT are trained from a general knowledge corpus (i.e. Wikipedia) and they differ in the size of the architecture. Small BERT has 4 encoder layers and 512 heads, whereas BERT has 12 encoder layers and 768 heads. BioClinical BERT is the same as BERT but it was fine-tuned in corpora from the medical domain including articles from PubMed. In practice, smaller models are more suitable for simpler tasks in order to avoid overfitting (i.e. no capacity to generalize for new data) such as Small BERT, while BERT is expected to have a better performance in more complex tasks. As BioClinical BERT has the same architecture as BERT but is trained with different data, it is expected to have a slight improvement with respect to BERT in those tasks which are from a technical and medical area.

While the models based on CQs had to be suited for each age-segment (training a particular f_{CQ2CoD} for each age-segment), the models based on ORs are transparent in the sense that a single approach can cope with all the age segments, thus making the f_{OR2CoD} versatile. Nevertheless, the final dual-input model is not only trained and tested with the whole VA data-set. An additional set of experiments was carried out: the input data for OR analysis has been divided into the different age groups (i.e. adult, child and neonate). Consequently, the amount of data provided for the OR2CoD approaches is significantly reduced for the fine-tuning process of these approaches, as is the number of target CoDs to be predicted. We also wonder if the proposed pre-trained models in OR2CoD will suffer more from a data amount decrease rather than from having less CoD to ascertain.

4.3. Models based on dual input

In order to combine both the OR and the CQs, we propose a dual-input approach. A model based on CQs (e.g. Tariff 2.0 or XGBoost) and a model based on the OR (e.g. XGBoost or BERT) mentioned, respectively, in Sections 4.1 and 4.2, are assembled to cope with the insights extracted from both the CQs and the OR. Fig. 2 depicts the architecture proposed.

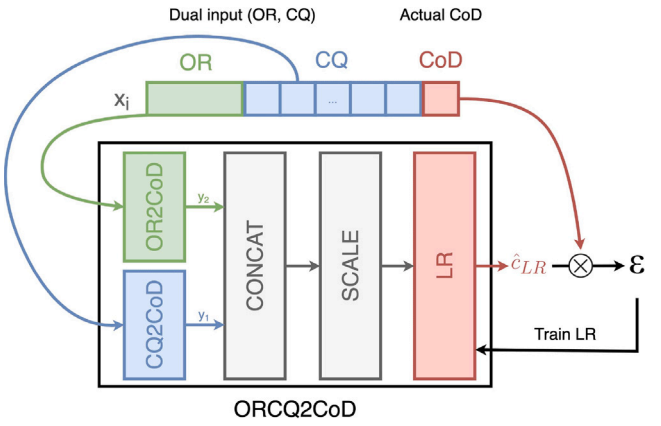

Fig. 2. Proposed ensemble model architecture having XGBoost for Closed Questions (CQ) treatment and BERT model for the Open Response (OR). The output of the Logistic Regression (LR) $c_{\hat{LR}}$ is the final prediction and is compared with the actual Cause of Death (CoD). Finally, the error (ϵ) is measured.

The dual input approach relies upon both f_{CQ2CoD} and f_{OR2CoD} to obtain, respectively, $\mathbf{y_1} \in [0,1]^{|CoD|}$ and $\mathbf{y_2} \in \mathbb{R}^{|CoD|}$, each of which determines the reliability of the available CoDs. Next, both weight-vectors are concatenated, leading to $\mathbf{y} = (\mathbf{y_1}, \mathbf{y_2}) \in [0,1]^{|CoD|} \times \mathbb{R}^{|CoD|}$. The output of each model ($\mathbf{y_1}$ and $\mathbf{y_2}$) has a slightly different meaning. The output given by the models based on CQs ($\mathbf{y_1}$) is probabilistic, values bound to $[0,1]$. By contrast, in the case of the models based on ORs the output, $\mathbf{y_2}$, can entail real values either positive or negative. In order to combine both outputs, an scaling operation is applied in order to adjust the meaning of each output. Note that, the concatenation operation is computed for each input instance. By contrast, the standardization is computed for each input-attribute.

The result of the transformation is, now, the input feature-vector of a simple logistic regression approach, and is denoted as $\mathbf{x}' = standardize(concat(\mathbf{y_1}, \mathbf{y_2}))$.

$$f_{CQOR2CoD} : \mathbb{R}^{2|CoD|} \longrightarrow \mathbb{R}^{|CoD|} \qquad (8)$$
$$\mathbf{x}' \longrightarrow f_{CQOR2CoD}(\mathbf{x}') = (y'_1, \ldots, y'_{|CoD|}) = \mathbf{y}'$$

The logistic regression is a classifier that learns a hyperplane for each of the CoD that tries to minimize the error by separating as much as possible the VAs that have the same CoD [29]. This hyperplane is inferred with a linear combination between the input and a set of weights (these are learnt in the training process). The capability of this classifier to combine the input is beneficial as it will automatically find the best combination of the input that maximizes the performance of CoD prediction during the training. The set of weights that are learnt is defined in (10) and the output will be a probability for each CoD computed from a linear combination between the input and the set of weights (i.e. α_i and β_i for the ith CoD). The main objective is to maximize the prediction accuracy and, in order to validate this method, the logistic regression will have at least the accuracy of the input model which has previously achieved the highest accuracy. As a result, it is expected that the output of both input models will need to be taken into account in order to increase the perform of using both of them separately.

In (9), the input ($\mathbf{x}'$) is made up of the concatenation and the standardization of the outputs of the transformer and the XGBoost model. The output of the logistic regression (y_{LR}) is the confidence assigned to each of the $|CoD|$ codes. This time, the final output will be $\hat{c}_{CQOR2CoD}$, that is, the predicted CoD. Training the logistic regression entails an optimization process in order to improve the performance in the CoD prediction.

Some preliminary experiments were carried out with a simple baseline approach that opted for the CoD with the highest probability in $(z_1, \ldots, z_{|CoD|}) \in \mathbb{R}^{|CoD|}$, with the rationale of what was done in (7), leading to lower evaluation scores. Thus, the need of a more complex way of combining the input is required and that is the reason why a logistic regression approach is employed in our approach as in (9). In addition, the logistic regression layer learns the best overall combination of the input that maximizes the performance.

$$\mathbf{x}' = standardize(concat(\mathbf{y_1}, \mathbf{y_2})) \qquad (9)$$
$$y_{LR}(\mathbf{x}') = (z_1, \ldots, z_{|CoD|}) \in \mathbb{R}^{|CoD|} \text{with} z_i \quad \text{as in (10)}$$
$$\hat{c}_{CQOR2CoD} = \arg\max_{i=1}^{|CoD|} z_i$$

The output of the logistic regression, $\mathbf{y}_{LR} = (y_1, \ldots, y_{|CoD|})$, is computed as in (10), where z_i represents a confidence weight of the input VA being associated with the ith CoD.

$$z_i = \vec{\alpha}_i \times \mathbf{x}' + \beta_i \quad \text{with} \qquad (10)$$
$$\vec{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{i|\mathbf{x}'|}) \in \mathbb{R}^{2|CoD|}$$

In the computation, the parameters $\vec{\alpha}_i$ are inferred in the training process of the logistic regression approach, which has a vector of weights ($\vec{\alpha}_i$) associated and learnt during the training process of the logistic regression. If the learnt weight for the k input, that is α_{ik}, is large, the output given for the ith cause of death (i.e. z_i) will be influenced by the value given by the k input. Extracting these weights leaves room for interpretability, as will be shown in Figs. 4 and 5, where each row corresponds to the learnt α_i vector and each column is the k input of the model.

5. Experimental results and discussion

In this section we offer and discuss the results given by the aforementioned methods. In Section 5.1, a comparison between XGBoost and Tariff 2.0 is made, having the CQs as input. In Section 5.2, instead, a XGBoost and a variety of transformer-based models are compared for the treatment of the OR. Finally, in Section 5.3, the results of the proposed ensemble model are shown.

5.1. Assessment of models based on closed questions

In this first experiment, the objective is to measure the performance of the WHO **standard** Tariff 2.0 for CoD prediction and compare it with the proposed XGBoost model.

In the training stage, several parameters were adjusted to optimize the performance of the resulting XGBoost model leading to the following values: maximum depth to 2, learning rate value to 0.25, sample type equal to 'weighted' and a grow policy as 'lossguide'. We have found that η has a great impact on the performance and that the optimal number of rounds (i.e. training iterations) is 75 for adults, 15 for children and 33 for neonates.

The per-class (per CoD) weighted averaged results of this experiment can be seen in Table 3. Note that these results are given only for the adult age group (and not for all the age groups) due to the fact that the OpenVA package including the Tariff 2.0 method only supports the adult age segment.

An inspection of Table 3 confirms that the results attained by Tariff 2.0 in our experiments are comparable to the results achieved in the antecedents [11], with an accuracy of 32.37%. Even more remarkable are, however, the results attained by XGBoost, outperforming the Tariff 2.0 in all the evaluation metrics. For example, the accuracy for the adult group attained by Tariff 2.0 is 37.39%, while by XGBoost it is 50.61%, leading to an improvement of more than 13% in terms of accuracy (the same applies to the F1 score metric).

Table 3
Assessment of the models based on CQs (denoted as CQ2CoD in Fig. 2): comparison between Tariff 2.0 and XGBoost for the adult age group by means of accuracy, precision, recall and F1 score.

Models based on CQs	Age group	Accuracy	Precision	Recall	F1 score
Tariff	Adult	37.39	39.05	37.39	35.59
XGBoost	Adult	50.61	49.13	50.61	50.22

In an attempt to ease reproducibility and, as a secondary contribution, with this article we have made available an R notebook[2] where the XGBoost and the Tariff 2.0 are compared for the adult group in two widely employed evaluation approaches: a 10-fold cross validation and a hold-out evaluation. The comparison between both models with the hold-out evaluation (i.e. splitting the data-set into train and test sets) is done with the data partitions described in Table 1.

With XGBoost clearly outperforming Tariff 2.0 and given that XGBoost can be implemented for all the age-groups, herein after, we continue our study with the XGBoost approach and do not restrict our evaluation to the adult age-group. Instead, we shall show the experimental results for all the age groups.

5.2. Assessment of models based on open response

In this section, we are concerned with the use of the OR to extract the CoD. We carried out a comparison between different approaches based on transformer architecture: XGBoost, Small BERT, BioClinical BERT and BERT. The goal is to determine if valuable information can be extracted from the OR and decide which model shows the best performance.

The three BERT-based models were fine-tuned with the training corpus by means of the Transformers python package [33] for sequence classification. In Fig. 3, a comparison of the performance of the three transformer-based models is given as the number of training epochs increases. The BERT model has the highest accuracy in all the epochs compared with the other two approaches, obtaining the highest score at the end of the training with an accuracy of 47.55% for the prediction of 48 different CoDs. Furthermore, the BioClinical BERT seems to behave like BERT with a lower performance but close until the fifth epoch where it begins to behave more like the Small BERT in terms of accuracy. Eventually, BERT (employing the OR), on its own, outperforms the results attained by Tariff 2.0 (employing the CQs).

In Table 4, weighted averaged CoD assessment for the 4 models: XGBoost as in [23], Small BERT, BERT and BioClinical BERT are shown by age-segment. While the CQs depend on the age-segment, the OR is a free text conveying an explanation and the format does not vary by age-segment, enabling, thus, the analysis of all the age groups in a versatile way. The results for the XGBoost with the OR reported by the antecedents [23] are compared with the transformer-based approaches explored in this work. We have not tested the XGBoost for the OR, as techniques such as word-embeddings or BoW reach high dimensionality representations which are not so appropriate for a XGBoost classifier for language modeling as the proposed transformer-based models.

In general, BERT seems to achieve the best results except for the adult set, where the accuracy and the recall are higher for the XGBoost model, and for neonate, where BioClinical BERT obtains the best score for precision and F1 score. Taking into account that this corpus contains health related terms, BioClinical BERT could be expected to get a higher score than the other two transformers-based models. Nonetheless, BERT surpasses BioClinical BERT in most cases and it can be due to the corpus used by BioClinical BERT, which is far more technical than the OR collected in the VAs. Even though the VA-GS corpus is framed in a medical area, the OR collects what the interviewee says and can be considered to be text of general knowledge and, consequently, non-technical.

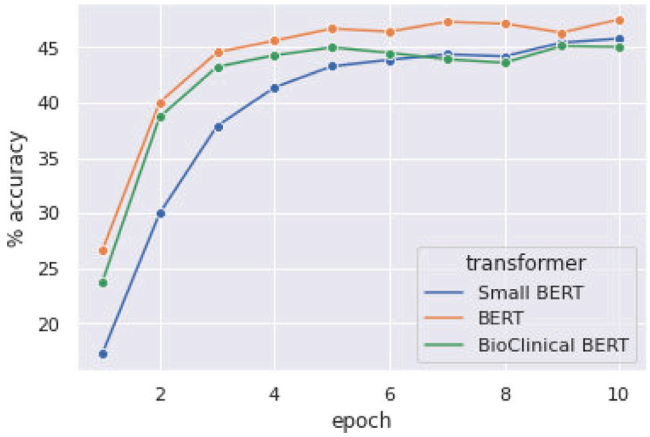

Fig. 3. Assessment of the models based on OR (denoted as OR2CoD in Fig. 2): comparison between three transformer-based models by means of accuracy on the test set for each of the training epochs for all the age groups.

We found that, overall (looking at the group denoted as 'All'), BERT is the best model to ascertain the cause of death with the OR in terms of accuracy and F1 but the difference with the XGBoost model for adults is small. Finally, as a conclusion of this experiment, it seems that valuable information can be extracted from the OR and we hypothesize that adding this information to these CQs could be useful for improving the performance when predicting the CoD.

5.3. Assessment of models based on dual input

In this final set of experiments we aim to combine the best two models obtained in Sections 5.1 and 5.2 in order to make use of both the OR and the CQs to predict the CoD. For the OR, the BERT model will be used while for the CQs, the XGBoost has proved to be the most suitable model for CoD prediction. Consequently, we wish to determine whether the OR adds valuable information or not. Indeed, this is one of the key points of this work. The results of these sets of experiments are shown in Table 5, given that the proposed ensemble models are based on the architecture represented in Fig. 2.

Table 5 reveals that the results obtained for all the age groups have been improved using both CQ and OR compared with just the use of the OR separately (as shown in Table 4). For instance, for all the age groups (i.e. denoted as 'All' in Table 5) the ensemble model attains a 56.20% in accuracy, while BERT achieves 47.55% in Table 4, having only the OR in the input. Thus, adding the CQs leverages the accuracy to 56.20%, resulting in a performance increase of around 8 points.

It is also remarkable that the performance attained by the ensemble model for the adult age group, which is 51.57% in accuracy,is nearly the same when it is compared to the 50.61% obtained when just using XGBoost with CQs as input in Table 3. This implies an increase of almost 1%. We hypothesize that the ensemble model benefits when having more data, as is the case of the 'All' age group. This shows a greater improvement compared with only using the OR for CoD prediction, even though there are more CoDs to predict, unlike the adult age group where there is less data and fewer CoDs.

[2] The software developed is made available with the user-name `IX-AVA` and password `Vaor2022?` at the following url: https://ixa2.si.ehu.eus/VAOpenResponse. Any use or any modification bound to the citation of this article.

Table 4
Assessment of the models based on OR (denoted as OR2CoD in Fig. 2): our experimental BERT-based results compared to the XGBoost in the antecedents [23] (the antecedents do not report the 'All' perspective). The results are presented per age group and measured with the weighted averaged CoD prediction accuracy, precision, recall and F1 score. The best accuracy and F1 results per age group are boldfaced.

Models based on OR	Age group	Accuracy	Precision	Recall	F1 score
XGBoost [23]	Adult	**45.60**	46.00	45.60	44.70
	Child	46.90	44.50	46.90	43.70
	Neonate	59.30	54.20	59.30	55.30
Small BERT	Adult	43.97	48.25	43.97	45.16
	Child	51.01	69.07	51.01	56.63
	Neonate	58.35	66.97	58.35	61.95
	All	45.81	51.04	45.81	47.42
BERT	Adult	45.48	48.57	45.48	**46.28**
	Child	**53.78**	69.21	53.78	**58.93**
	Neonate	**61.80**	71.80	61.80	64.79
	All	**47.55**	51.66	47.55	**48.85**
BioClinical BERT	Adult	43.63	47.18	43.63	44.62
	Child	51.26	60.35	51.26	54.18
	Neonate	60.74	75.81	60.74	**66.49**
	All	45.05	47.42	45.05	45.74

Table 5
Assessment of the model based on dual input (denoted as ORCQ2CoD in Fig. 2): the CQs are handled by XGBoost, the OR handled by BERT and the final output given by the logistic regression. The results are presented per age group and measured with the weighted averaged CoD prediction accuracy, precision, recall and F1 score.

Model based on dual input	Age group	Accuracy	Precision	Recall	F1 score
XGBoost+BERT+Logistic Regression	Adult	51.57	50.89	51.57	50.89
	Child	54.29	53.38	54.29	52.77
	Neonate	70.82	68.87	70.82	69.17
	All	56.20	56.17	53.87	53.87

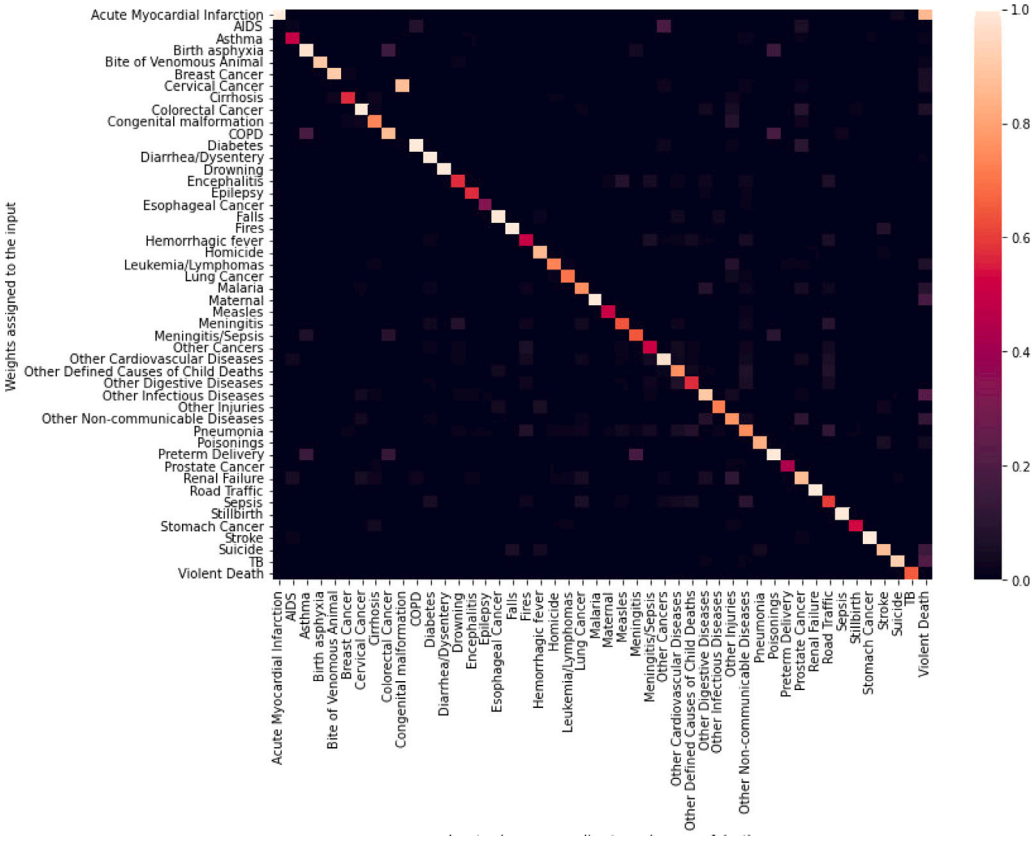

Fig. 4. Heat-map of the weights learned in the last layer of the ensemble model of Fig. 2 by logistic regression for the input given by the XGBoost model scaled between 0 and 1. The y axis represents the weights learnt for the final output, while the x axis indicates to which of the XGBoost outputs (i.e. probability given by XGBoost for a particular CoD) that weight corresponds.

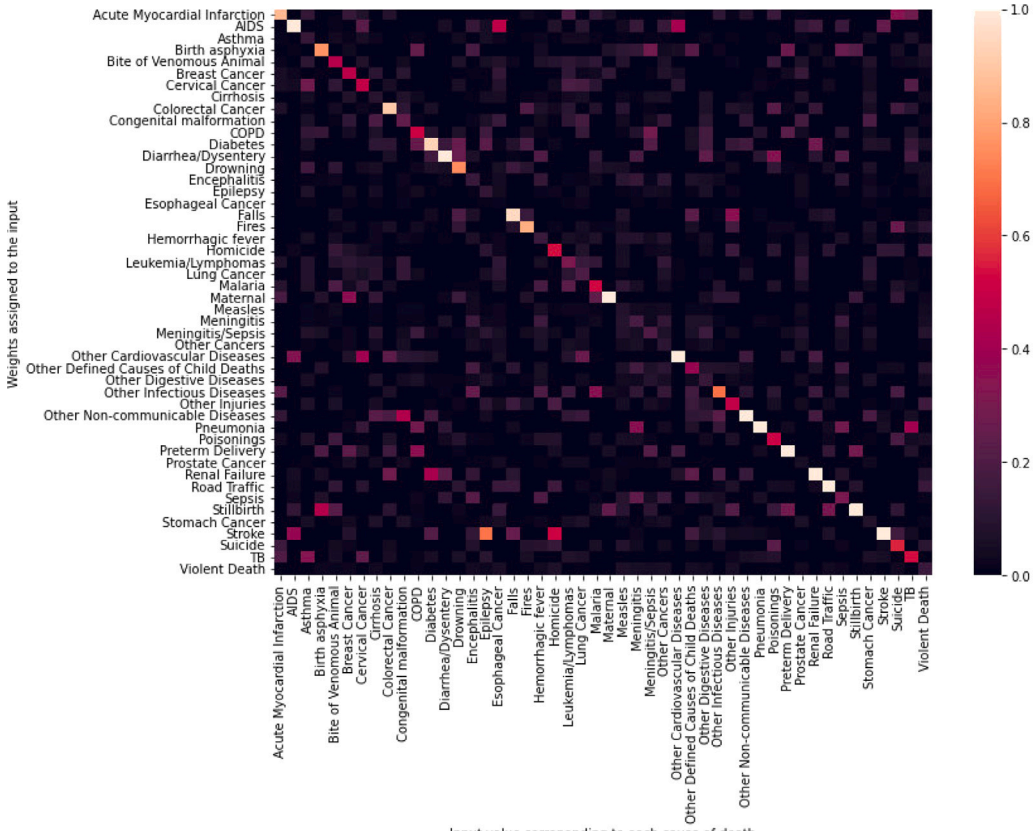

Fig. 5. Heat-map of the weights learned in the last layer of the ensemble model of Fig. 2 by logistic regression for the input given by the BERT model scaled between 0 and 1. The y axis are the weights learnt for the final output while the x axis indicates to which of the BERT outputs (i.e. value given by BERT for a particular CoD) corresponds that weight.

An interesting **research question** arising was to know the extent to which the logistic regression relied on two inputs, OR and CQ, in making decisions for the ensemble model. We have inspected the parameters involved in the combination, see expression, (10) and graphically depicted them in heat-maps as shown in Figs. 4 and 5 to study the credibility given by the logistic regression to the values provided in the input by XGBoost and BERT in predicting each CoD. Each row represents the set of weights assigned to each CoD proposed by XGBoost or BERT, which indicates how reliable the CoD value given by the models is. The bigger the weight, the more relevant the input is for the logistic regression when predicting the CoD.

Taking Figs. 4 and 5 into account, we can derive that the ensemble relies on both the BERT model and XGBoost model, since in both cases, several of the CoDs of the main diagonal have been assigned a large weight. For example, for the output corresponding to 'Suicide' the logistic regression was relying in both BERT and XGBoost. However, for BERT, it was not only relying in the output given for 'Suicide', it was also relying in the output given for 'Homicide'. In Fig. 1, we can see an example of an OR, whose CoD is 'Suicide', but the BERT model has given the highest value to 'Homicide'. Because this seems to be a common mistake, the logistic regression relies on both values, hence, the logistic regression is also able to reduce the impact of the common mistakes that both input models can make.

As a conclusion, first, if we compare the predictive capability of the CQ, we find that XGBoost is superior to Tariff 2.0 for every age group. Additionally, simple and imperfect though the OR might seem, it has proven able to convey competitive predictive capabilities and is superior to Tariff 2.0. Finally, when the dual input is available, marginal improvements are attained over the XGBoost, though significant improvements with respect to Tariff 2.0. In any case, the OR gains importance over the CQ in the ensemble approach.

Furthermore, although the proposed ensemble model has achieved a superior classification performance with respect to previous approaches, there is still room for improvement compared to the results obtained on related death certification tasks [34,35]. In addition, it can be of interest to conduct a more detailed analysis by considering a confidence score. This score would separate predictions with a high probability of corresponding to a certain CoD with those predictions with a low probability that may lead to prediction errors. This could help data producers know when human expertise may be necessary to handle cases whose CoD is difficult to estimate.

Finally, let us raise a general concern inherent in the data. Supervised learning approaches rest on the quality of the annotations in the gold-standard and, hence, are bound to the underlying annotation errors. That is, provided that the data was erroneously labeled, the models would be inferred from noisy samples. McCormick et al. [5] show that the labels can differ between different expert physicians, especially when their training exposure and speciality vary.

6. Conclusions

This work considered the use of Open Response available in Verbal Autopsies in the estimation of the cause of death, instead of limiting the estimation to the information available in Closed Questions (as does Tariff 2.0, a model made available within the WHO 2016 instrument). State-of-the-art Natural Language Processing techniques were applied and proven useful to extract valuable information.

Our approach outperformed Tariff 2.0. In fact, experimental results showed that the OR brings significant information and results in a valuable source of information for CoD ascertainment. This finding opens pathways towards the simplification of the Verbal Autopsy interview.

Data availability and quality is stressed as being a main limitation of supervised algorithms and, thus, we would encourage the community

to release these data to promote research in this field. This is a complex multi-class classification task and the models developed in this work, especially BERT, would benefit from further data.

Ethical

This article does not contain any studies with human participants or animals performed by any of the authors.

Declaration of competing interest

The authors declare that there is no conflict of interest.

Acknowledgments

This work was partially funded by the Spanish Ministry of Science and Innovation (DOTT-HEALTH/PAT-MED PID2019-106942RB-C31); by both Antidote PCI2020-120717-2 and Lotu TED2021-130398B-C22 funded by the MCIN/AEI /10.13039/501100011033 and by the European Union NextGenerationEU/ PRTR; by the Basque Government (IXA IT-1570-22); and by Misiones Euskampus 2.0 (EXTEPA).

References

[1] WHO. Civil registration: why counting births and deaths is important. 2014, https://www.who.int/news-room/fact-sheets/detail/civil-registration-why-counting-births-and-deaths-is-important. (Accessed 19 April 2022).
[2] Byass P, Hussain-Alkhateeb L, D'Ambruoso L, Clark S, Davies J, Fottrell E, et al. An integrated approach to processing WHO-2016 verbal autopsy data: the InterVA-5 model. BMC Med 2019;17(1):1–12.
[3] Chandramohan D, Fottrell E, Leitao J, Nichols E, Clark SJ, Alsokhn C, et al. Estimating causes of death where there is no medical certification: evolution and state of the art of verbal autopsy. Glob Health Action 2021;14(sup1):1982486, PMID: 35377290.
[4] Murray CJ, Lopez AD, Black R, Ahuja R, Ali SM, Baqui A, et al. Population Health Metrics Research Consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. Popul Health Metr 2011;9(1):27.
[5] McCormick TH, Li ZR, Calvert C, Crampin AC, Kahn K, Clark SJ. Probabilistic cause-of-death assignment using verbal autopsies. J Amer Statist Assoc 2016;111(515):1036–49.
[6] 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 Med 2015;13(1):291.
[7] Flaxman AD, Joseph JC, Murray CJ, Riley ID, Lopez AD. Performance of InSilicoVA for assigning causes of death to verbal autopsies: multisite validation study using clinical diagnostic gold standards. BMC Med 2018;16(1):1–11.
[8] Jha P, Gajalakshmi V, Gupta PC, Kumar R, Mony P, Dhingra N, et al. Prospective study of one million deaths in India: rationale, design, and validation results. PLoS Med 2006;3(2):e18.
[9] Danso S, Atwell E, Johnson O, H A, Soremekun S, Edmond K, et al. A semantically annotated verbal autopsy corpus for automatic analysis of cause of death. ICAME J Int Comput Arch Mod Mediev Engl 2013;37:37–69.
[10] Ganapathy SS, Yi Yi K, Omar MA, Anuar MFM, Jeevananthan C, Rao C. Validation of verbal autopsy: determination of cause of deaths in Malaysia 2013. BMC Public Health 2017;17:1–8.
[11] Li ZR, McCormick TH, Clark SJ. Verbal autopsy analysis using openVA. 2018.
[12] Li R, Thomas J, Choi E, McCormick T, Clark S. The openVA toolkit for verbal autopsies. 2021.
[13] Rish I, et al. An empirical study of the naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3. 2001, p. 41–6.
[14] Flaxman AD, Joseph JC, Murray CJL, Riley ID, Lopez AD. Performance of InSilicoVA for assigning causes of death to verbal autopsies: multisite validation study using clinical diagnostic gold standards. BMC Med 2018;16(1):56.
[15] Danso S, Atwell E, Johnson O. A comparative study of machine learning methods for verbal autopsy text classification. IJCSI Int J Comput Sci Issues 2013.
[16] Noble WS. What is a support vector machine? Nature Biotechnol 2006;24(12):1565–7.
[17] Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005;12(3):296–8.
[18] Ramos J, et al. Using TF-IDF to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning, vol. 242, no. 1. 2003, p. 29–48.
[19] Yan Z, Jeblee S, Hirst G. Can character embeddings improve cause-of-death classification for verbal autopsy narratives? In: Proceedings of the 18th BioNLP workshop and shared task. 2019, p. 234–9.
[20] Moran KR, Turner EL, Dunson D, Herring AH. Bayesian hierarchical factor regression models to infer cause of death from verbal autopsy data. J R Stat Soc Ser C Appl Stat 2021;70(3):532–57.
[21] Li ZR, McComick TH, Clark SJ. Using Bayesian latent Gaussian graphical models to infer symptom associations in verbal autopsies. Bayesian Anal 2020;15(3):781.
[22] Reynolds DA. Gaussian mixture models. Encycl Biom 2009;741(659–663).
[23] Blanco A, Pérez A, Casillas A, Cobos D. Extracting cause of death from verbal autopsy with deep learning interpretable methods. IEEE J Biomed Health Inf 2020;25(4):1315–25.
[24] 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 Med 2018;15(1):e1002486.
[25] Gao Y, Liu W, Lombardi F. Design and implementation of an approximate softmax layer for deep neural networks. In: 2020 IEEE international symposium on circuits and systems. IEEE; 2020, p. 1–5.
[26] Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. Xgboost: eXtreme gradient boosting 1 (4). 2015, p. 1–4, R Package Version 0.4-2.
[27] Shao Y, Taylor S, Marshall N, Morioka C, Zeng-Treitler Q. Clinical text classification with word embedding features vs. bag-of-words features. In: 2018 IEEE international conference on big data. 2018, p. 2874–8.
[28] Mikolov T, Karafiát M, Burget L, Cernockỳ J, Khudanpur S. Recurrent neural network based language model. In: Interspeech, vol. 2, no. 3. Makuhari; 2010, p. 1045–8.
[29] Kleinbaum DG, Dietz K, Gail M, Klein M, Klein M. Logistic regression. Springer; 2002.
[30] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst 2017;30.
[31] Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 conference of the north American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (Long and short papers). Association for Computational Linguistics; 2019, p. 4171–86. http://dx.doi.org/10.18653/v1/n19-1423.
[32] Alsentzer E, Murphy J, Boag W, Weng W-H, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd clinical natural language processing workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019, p. 72–8.
[33] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations. Association for Computational Linguistics; 2020, p. 38–45, Online.
[34] Popescu MH, Roitero K, Travasci S, Della Mea V. Automatic Assignment of ICD-10 Codes to Diagnostic Texts using Transformers Based Techniques. In: 2021 IEEE 9th international conference on healthcare informatics. 2021, p. 188–92.
[35] Falissard L, Morgand C, Ghosn W, Imbaud C, Bounebache K, Rey G, et al. Neural translation and automated recognition of ICD-10 medical entities from natural language: Model development and performance assessment. JMIR Med Inform 2022;10(4):e26353.