A Deep Learning Method for ICD-10 Coding of Free-Text Death Certificates

Francisco Duarte[1](✉), Bruno Martins[1], Cátia Sousa Pinto[2], and Mário J. Silva[1]

[1] INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
{francisco.ribeiro.duarte,bruno.g.martins,
mario.gaspar.silva}@tecnico.ulisboa.pt
[2] Direção-Geral da Saúde, Lisbon, Portugal
catiasousapinto@dgs.min-saude.pt

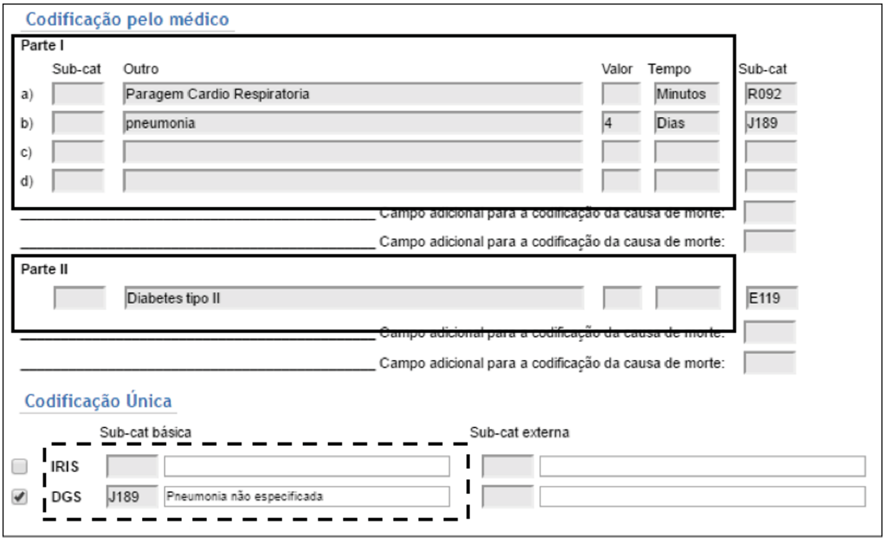
Abstract. The assignment of disease codes to clinical texts has a wide range of applications, including epidemiological studies or disease surveillance. We address the task of automatically assigning the ICD-10 codes for the underlying cause of death, from the free-text descriptions included in death certificates obtained from the Portuguese Ministry of Health. We specifically propose to leverage a deep neural network based on a two-level hierarchy of recurrent nodes together with attention mechanisms. The first level uses recurrent nodes for modeling the sequences of words given in individual fields of the death certificates, together with attention to weight the contribution of each word, producing intermediate representations for the contents of each field. The second level uses recurrent nodes to model a sequence of fields, using the representations produced by the first level and also leveraging attention in order to weight the contributions of the different fields. The paper reports on experiments with a dataset of 115,406 death certificates, presenting the results of an evaluation of the predictive accuracy of the proposed method, for different ICD-10 levels (i.e., chapter, block, or full code) and for particular causes of death. We also discuss how the neural attention mechanisms can help in interpreting the classification results.

Keywords: Classification of death certificates · Clinical text mining · Deep learning · Natural language processing · Artificial intelligence in medicine

1 Introduction

The systematic collection of high-quality mortality data is essential in the context of monitoring a population's health, also serving as a basis for mortality and epidemiologic studies. For this and other legal purposes, doctors write death certificates, i.e. reports including the deceased personal data and textual descriptions for the cause of death, as well as any contributing conditions or injuries. In Portugal, doctors are now submitting death certificates in electronic format, using a national Death Certificate Information System (SICO [1]) for data collection and registry purposes. The analysis and classification of causes of death are

E. Oliveira et al. (Eds.): EPIA 2017, LNAI 10423, pp. 137–149, 2017.
DOI: 10.1007/978-3-319-65340-2_12



Codificação pelo médico

Parte I

	Sub-cat	Outro	Valor	Tempo	Sub-cat
a)		Paragem Cardio Respiratoria		Minutos	R092
b)		pneumonia	4	Dias	J189
c)					
d)					

Campo adicional para a codificação da causa de morte:

Campo adicional para a codificação da causa de morte:

Parte II

	Diabetes tipo II			E119

Campo adicional para a codificação da causa de morte:

Campo adicional para a codificação da causa de morte:

Codificação Única

		Sub-cat básica		Sub-cat externa	
□	IRIS				
☑	DGS	J189	Pneumonia não especificada		

Fig. 1. The form in the SICO platform that is used for coding the death certificates.

based on Revision 10 of the International Classification of Diseases, ICD-10, the standard medical classification list developed and reviewed by the World Health Organization. However, the assignment of ICD codes to the death certificates provided by doctors is is still made manually by mortality coders with specific expertise, based on the free-text descriptions included in the death certificates.

Figure 1 presents a screenshot of the SICO form used by mortality coders in Portugal to assign ICD-10 codes to death certificates. The form has two parts (delimited by the solid lines). Part I has four rows of text, labelled *(a)*, *(b)*, *(c)* and *(d)*, for reporting a chain of events leading directly to death. The underlying causes of death should be provided in the lowest line(s) and the immediate cause of death in the first one. Part II is filled-in only if necessary for reporting other significant diseases, conditions or injuries that contributed to death, but are not part of the main causal sequence leading to it. After the manual review of the data, the mortality coder should assign the corresponding ICD-10 code, in the box shown under the dashed line of Fig. 1.

The manual coding of the free-text contents in death certificates is a challenging, expensive, and time consuming task [2], which slows down the dissemination of mortality statistics and prevents real time surveillance. However, we believe that the large number of certificates that have been manually coded in the past can be used to support supervised machine learning of models for automatically assigning codes to the certificates. Automated approaches can be used to speed-up the process of publishing mortality statistics by quickly producing results that can latter be revised through manual coding. When integrated into existing platforms, automated approaches can also provide suggestions to assist the manual coders. If sufficiently accurate, automatic coding also has the potential to significantly reduce the costs with human experts, and to increase coding consistency.

Several previous studies have already addressed the ICD coding of free-text death certificates [3–6]. Recently, increasing attention has been given to this problem due to the organization of CLEF eHealth clinical information extraction tasks in 2016 and 2017, which involved large-scale datasets prepared from French and English death certificates [7,8]. However, previously published methods are still relatively simple in comparison to the current state-of-the-art in other text classification problems, either leveraging dictionary projection methods or supervised machine learning with linear models and manual feature engineering.

In this paper, we propose to leverage a neural network based on a two-level hierarchy of recurrent nodes together with attention mechanisms, inspired on previous work by Yang et al. [9]. The first level of the model uses Gated Recurrent Units (GRUs) [10] for modeling the sequences of words given in individual fields of the death certificates, together with attention to weight the contribution of each word, producing intermediate representations for the contents of each field. The second level uses GRUs to model a sequence of fields, using the representations produced by the first level and also leveraging attention in order to weight the contributions of the different fields. The representations produced by the second level are passed to feed-forward nodes, which leverage a softmax activation to predict the most likely ICD-10 codes. The entire model can be trained end-to-end from a set of coded death certificates, leveraging the back-propagation algorithm in conjunction with the Adam optimization method [11,12].

The paper reports on experiments with a dataset of 115,406 death certificates from the years of 2013 up to 2015, through which we evaluated the predictive accuracy of the proposed method. The available data was randomly split into two subsets (i.e., 75% for model training and 25% for testing), and we measured results in terms of classification accuracy, as well as macro-averaged precision, recall, and F1-scores. Given the hierarchical organization of ICD-10 (i.e., the codes are organized hierarchically into chapters, blocks and full codes), we also measured results according to different levels of specialization.

Our best model achieved an accuracy of 86%, 78%, and 75%, respectively when considering ICD-10 chapters (i.e., a total of 22 different classes appearing in our dataset), blocks (i.e., 697 different classes) and full codes (i.e., 1,674 different classes). Our full model also achieved F1-scores of 96% and 90%, respectively in terms of correctly identifying causes of mortality related to ICD-10 Chapters II (i.e., neoplasms) and IX (i.e., diseases of the circulatory system), that together represent 58.7% of the death causes in the dataset. We argue that the obtained results indicate that automatic approaches leveraging supervised machine learning can indeed contribute to a faster processing of death certificates, given the relatively low classification error. Moreover, although our experiments failed to show that neural attention mechanisms lead to an increased performance, these methods can offer much needed model interpretability, by allowing us to see which parts of the input are attended to when making predictions.

2 Related Work

Various previous studies have addressed automatic ICD-10 coding. For instance Koopman et al. described the use of Support Vector Machines (SVMs) for identifying cancer related causes of death in natural language death certificates [5]. The textual contents were encoded as binary feature vectors (i.e., vectors encoding the presence of terms, term n-grams, and SNOMED CT concepts recognized by a clinical natural language processing system named Medtex), and these representations were used as features to train a two-level hierarchy of SVM models: the first level was a binary classifier for identifying the presence of cancer, and the second level consisted of a set of classifiers (i.e., one for each cancer type) for identifying the type of cancer according to the ICD-10 classification system (i.e., according to 85 different ICD-10 blocks, of which 20 instances corresponded to 85% of all cases). The system was highly effective at identifying cancer as the underlying cause of death (i.e., a macro-averaged F1-score of 0.94 for the first level classifier). It was also effective at determining the type of common cancers (i.e., a macro-averaged F1-score of 0.7), although rare cancers, for which there was little training data, were difficult to classify accurately (i.e., a macro-averaged F1-score of 0.12). The principal factors influencing performance were the amount of training data and certain ambiguous cases (e.g., cancers in the stomach region).

In a separate study, Koopman et al. described machine learning and rule-based methods to automatically classify death certificates according to four high impact diseases of interest, namely diabetes, influenza, pneumonia and HIV [6]. The rule-based method leveraged sets of keyword-matching rules, while the machine learning method was again based on SVM classifiers, using binary feature vectors (i.e., presence of terms, term n-grams, and SNOMED CT concepts recognized by Medtex) for encoding the texts. In the case of the machine learning approach, a separate model was trained for each of the four diseases of interest, and the authors also experimented with more fine-grained classifiers trained for each of the relevant ICD-10 blocks. An empirical evaluation was conducted using 340,142 certificates (i.e., 80% for model training and 20% for testing) covering deaths from 2000–2007 in New South Wales, Australia. The results showed that the classification of diabetes, influenza, pneumonia and HIV was highly accurate (i.e., a macro-averaged F1-score of 0.95 for the rule-based method, and 0.94 when using machine learning). More fine-grained ICD-10 classification had nonetheless a more variable effectiveness, with less accurate classifications for blocks with little training data available, although results were still high (i.e., a macro-averaged F1-score of 0.80, when discriminating over 9 different ICD-10 blocks). The error analysis revealed that word variations (e.g., *pneumonitis* or *pneumonic* as variants for *pneumonia*) as well as certain word combinations adversely affected classification. In addition, anomalies in the ground truth likely led to an underestimation of the effectiveness (i.e., the authors observed some class confusions, e.g. in ICD blocks E10 versus E11).

Mujtaba et al. tested different text classification methods in the task of coding death certificates with nine possible ICD-10 codes [4], aiming to assist pathologists in determining causes of death based on autopsy findings. The dataset used in these experiments was composed of 2200 autopsy reports obtained from one of the largest hospitals in Kuala Lumpur, and the methods under study involved different feature selection schemes, and also five different learning algorithms. Random forests and J48 decision models, parametrized using expert-driven feature selection and leveraging a feature subset size of 30, yielded the best results (e.g., approximately 90% in terms of the macro-averaged F1-score).

Lavergne et al. described a large-scale dataset prepared from French death certificates, suitable to the application of machine learning methods for ICD-10 coding [8]. The dataset comprised a total of 93,694 death certificates referring to 3,457 unique ICD-10 codes, and it was made available for international shared tasks organized in the context of CLEF. The 2016 edition of the CLEF eHealth shared task on ICD-10 coding attracted five participating teams, which presented systems relying either on dictionary linking or statistical machine learning [7]. The shared task was defined at the level of each statement (i.e., lines varying from 1 to 30 words, with outliers at 120 words and with the most frequent length at 2 tokens) in a death certificate, and statements could be associated with zero, one or more ICD-10 codes. The best-performing system achieved a micro-averaged F1-score (i.e., harmonic mean of precision and recall weighted by the class size) of 0.848, leveraging dictionaries built from the shared task data.

Leveraging the CLEF eHealth dataset, Zweigenbaum and Lavergne. presented hybrid methods for ICD-10 coding of death certificates [3], combining dictionary linking with supervised machine learning (i.e., an SVM classifier leveraging tokens, character tri-grams, and the year of the certificate as features). The best hybrid model corresponded to the union of the results produced by the dictionary- and learning-based methods, outperforming the best system at the 2016 edition of the CLEF eHealth shared task with a micro-averaged F1-score of 0.8586.

3 The Proposed Approach

We propose a neural network model for assigning ICD-10 codes to free-text death certificates, taking inspiration on previous work by Yang et al. [9]. Considering the SICO platform from the Portuguese Ministry of Health's Directorate-General of Health (DGS), illustrated on Fig. 1, we modeled the coding task as follows: given different textual strings encoding events leading to death, an automated system should output the ICD-10 code corresponding to the underlying cause of death. Figure 2 presents the neural network architecture, which is briefly explained next. For an in-depth introduction to deep neural networks for natural language processing, the reader can refer to the tutorial by Goldberg [13].

Noting that the certificates can be seen as having a hierarchical structure (i.e., words form different fields, and the fields from Parts I and II, as shown in Fig. 1, form the certificate), our model first builds representations of individual fields, and then aggregates those into a representation for the certificate. Both hierarchical levels are illustrated in Fig. 2, with the word-level part of the model

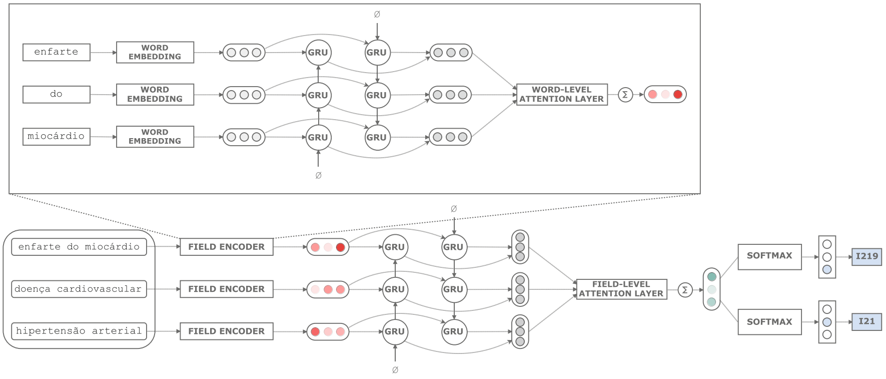


Fig. 2. The proposed neural architecture (with 3 fields instead of 6, for illustration).

shown in the box at the top. A recurrent neural network node known as a Gated Recurrent Unit (GRU) is used at both levels to build the representations, and we specifically considered bi-directional GRUs [10]. Notice that the GRUs in the first level of the model leverage word embeddings as input, whereas the second level uses as input the field representations generated at the first level.

GRUs model sequential data by having a recurrent hidden state whose activation at each time is dependent on that of the previous time. A GRU computes the next hidden state h_t given a previous hidden state h_{t-1} and the current input x_t using two gates (i.e., a reset gate r_t and an update gate z_t), that control how the information is updated, as shown in Eq. 1. The update gate (Eq. 2) determines how much past information is kept and how much new information is added, while the reset gate (Eq. 4) is responsible for how much the past state contributes to the candidate state. In Eqs. 1 to 4, $\tilde{h}_t$ stands for the current new state, W is the parameter matrix for the actual state, U is the parameter matrix for the previous state, and b is a bias vector.

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{1}$$

$$z_t = \sigma\big(W_z x_t + U_z h_{t-1} + b_z\big) \tag{2}$$

$$\tilde{h}_t = \tanh\big(W_h x_t + r_t \odot (U_h h_{t-1} + b_h)\big) \tag{3}$$

$$r_t = \sigma\big(W_r x_t + U_r h_{t-1} + b_r\big) \tag{4}$$

Bi-directional GRUs perceive the context of each input in a sequence by outlining the information from both directions. Concatenating the output of processing a sequence forward $\overrightarrow{h}_{it}$ and backwards $\overleftarrow{h}_{it}$ grants a summary of the information around each position, $h_{it} = [\overrightarrow{h}_{it}, \overleftarrow{h}_{it}]$.

Since the different words and fields can be differently informative in specific contexts, the model also includes two levels of attention mechanisms (i.e., one at the word level and one at the field level), that let the model pay more or less attention to individual words/fields when constructing representations.

For instance, in the case of the word-level part of the network, the outputs h_{it} of the bi-directional GRU encoder are provided to a feed-forward node (Eq. 5), resulting in vectors u_{it} representing words in the input. A normalized importance α_{it} (i.e., the attention weights) is calculated as shown in Eq. 6, using a context vector u_w that is randomly initialized. The importance weights in α_{it} are then summed over the whole sequence, as shown in Eq. 7.

$$u_{it} = \tanh\left(W_w h_{it} + b_w\right) \tag{5}$$

$$\alpha_{it} = \frac{\exp(u_{it}^T u_w)}{\sum_t \exp(u_{it}^T u_w)} \tag{6}$$

$$s_i = \sum_t \alpha_{it} h_{it} \tag{7}$$

The vector s_i from Eq. 7 is finally taken as the representation of the input. The part of the network that processes the sequence of fields similarly makes use of bi-directional GRUs with an attention mechanism, taking as input the representations produced for each field, as shown in Fig. 2.

The representation resulting from the different fields is finally passed to two feed-forward nodes with softmax activations, which respectively attempt to predict the corresponding ICD-10 block and the ICD-10 full-code. The entire model is trained end-to-end from a set of coded death certificates, leveraging the back-propagation algorithm [11] in conjunction with the Adam method [12] for optimizing loss functions corresponding to the categorical cross-entropy (i.e., we combine loss functions computed from ICD-10 blocks and full-codes, respectively with weights 0.5 and 1.0). The idea of leveraging two separate outputs relates to the large number of ICD-10 full-codes that are sparsely used. We expect that information on ICD-10 blocks can be used to better inform model training.

4 Experimental Evaluation

This section describes the experimental evaluation of the proposed method. We first present a statistical characterization of the dataset that supported our tests, together with the considered experimental methodology. Then, Subsect. 4.2 presents and discusses the obtained results, also giving illustrative examples.

4.1 Dataset and Experimental Methodology

The dataset used in our experiments consisted of death certificates emitted from the years of 2013 to 2015, collected from the SICO platform. We excluded all instances involving a supplemental autopsy report, mostly corresponding to accidents, suicides, or homicides. Table 1 presents characterization statistics.

Figure 1 already presented the general layout of the SICO online platform that is currently being used for manually coding the death certificates. For each certificate, we use the textual contents of fields labeled from *(a)* to *(d)* in Part

Table 1. Statistical characterization of the dataset used in our experiments.

Number of distinct ICD-10 codes	1,674
Number of distinct ICD-10 blocks	697
Number of distinct ICD-10 chapters	22
Number of certificates	115,406
Average number of fields with textual data	2.3
Average number of words per field	7.5
Maximum number of words per field	71
Vocabulary size	16,778

I, as well as the contents from Part II, in each case concatenating the strings labeled as *Outro*, *Valor* and *Tempo* (i.e., the fields named *Valor* and *Tempo* can be used to encode the approximated interval between the onset of the respective condition and the date of death, which can be relevant in cases like a stroke that occurred much before the time of death). Thus, each instance in the dataset consists of 6 different strings (i.e., we noticed that the field from Part II often contained two sentences), some of them possibly empty, padded with special symbols to encode the beginning/termination of the textual contents, together with the ICD-10 code corresponding to the main cause of death.

The available data was split into two subsets, with 75% (86,554 death certificates) for model training and 25% (28,852 certificates) for testing. Table 3 presents the distribution of the number of instances associated to each ICD-10 chapter (i.e., the column named *percentage* gives the fraction of instances, in the training plus the testing splits, corresponding to each chapter). Notice that some ICD-10 chapters have no instances in our dataset, given that the corresponding health problems are seldom related to death (i.e., Chapter VII, corresponding to diseases of the eye and adnexa), or are instead related to external causes that require an autopsy report (e.g., Chapter XIX, corresponding to injury, poisoning and certain other consequences of external causes).

All experiments relied on the keras[1] deep learning library. The word embedding layer in the first level considered a dimensionality of 50, and the output of the GRU layers had a dimensionality of 25. Model training was made in batches of 32 instances, using the Adam optimization algorithm [12] with default parameters. Model training considered a stopping criteria based on the training loss, finishing when the difference between epochs was less than 10^{-6}.

For accessing prediction quality, we measured the classification accuracy over the test split, as well the macro-averaged precision, recall and F1-scores (i.e., macro-averages assign an equal importance to each class, thus providing useful information in the case of datasets with a highly unbalanced class distribution).

[1] http://keras.io.

Given the hierarchical organization of ICD-10, we also measured results according to different levels of specialization (i.e., ICD-10 chapters, blocks, and full codes).

4.2 Experimental Results

Our experiments compared three different neural architectures: (i) a hierarchical model with two levels of GRUs but without the attention mechanisms, thus using the hidden states produced at the edges of the sequences in order to build the representations, an also considering only a single output node for the full ICD-10 code; (ii) a hierarchical attention model that also considers only a single output; (iii) the full model with two output nodes, as described in Sect. 3. Models (i) and (ii) correspond to variations were some of the components were removed.

Table 2 presents the results, and Table 3 further details the results obtained with Model (iii), by showing evaluation scores for each individual ICD-10 chapter. The best values is terms of accuracy were actually obtained with the simpler model, corresponding to 86%, 78%, and 75%, respectively when considering ICD-10 chapters, blocks and full-codes. To further access the overall performance of our method, we also computed the Mean Reciprocal Rank (MRR) of the correct class, when sorting classes according to the probability assigned prior to performing the softmax operation. Model (iii) has a MRR of 0.795 when assigning full codes, 0.830 for blocks, and 0.899 for chapters.

ICD Chapters II (i.e., neoplasms) and IV (i.e., diseases of the circulatory system) correspond to the most common causes of death in our dataset and, together, they represent approximately 58.1% of the instances. Table 4 further details the results obtained by Model (iii) in these two important chapters. We can also see that deaths with underlying cause in Chapter XVIII (i.e., symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified) were predicted with high effectiveness (i.e., an F1-score of 93.925%).

Table 2. Performance metrics for different variants of the neural model.

	ICD-10 level	Accuracy	Macro-averages		
			Precision	Recall	F1-score
Hierarchical model	Chapter	**86.417**	60.200	57.893	58.781
	Block	**78.459**	**35.786**	**32.824**	**32.892**
	Full code	**74.567**	**25.550**	**24.727**	**23.920**
+ attention mechanisms	Chapter	85.297	59.133	55.069	56.319
	Block	76.314	30.473	28.642	28.579
	Full code	72.480	20.760	20.417	19.471
+ two outputs	Chapter	86.372	**73.498**	**69.614**	**71.031**
	Block	78.171	33.919	31.614	31.658
	Full code	73.981	23.360	23.048	22.057

Table 3. Number of instances and obtained results for each of the ICD-10 chapters.

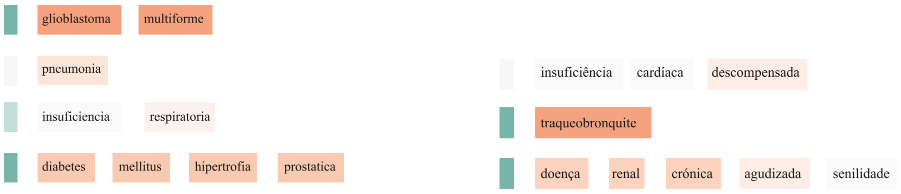
Chapter	Occurences		Percentage	Precision	Recall	F1-score
	Train	Test				
I	1, 952	642	2.248	65.994	66.199	66.096
II	23, 971	7, 921	27.634	95.608	95.922	95.765
III	418	126	0.471	36.947	32.540	34.599
IV	4, 836	1, 567	5.453	75.925	66.816	71.079
V	2, 386	849	2.803	76.162	75.265	75.711
VI	3, 033	1, 053	3.540	84.747	76.828	80.139
VII	0	0	0.000	—	—	—
VIII	8	1	0.008	100.000	100.000	100.000
IX	26, 773	9, 036	31.028	89.982	89.464	89.722
X	11, 109	3, 675	12.810	80.273	86.259	83.158
XI	3, 970	1, 334	4.596	78.012	80.585	79.277
XII	112	36	0.128	25.000	11.111	15.385
XIII	390	112	0.435	50.980	46.429	48.598
XIV	2, 695	905	3.119	64.870	69.171	66.952
XV	0	0	0.000	—	—	—
XVI	10	1	0.003	100.000	100.000	100.000
XVII	105	34	0.120	68.421	38.235	49.057
XVIII	3, 725	1, 166	4.238	90.778	97.084	93.925
XIX	0	0	0.000	—	—	—
XX	1, 171	394	1.356	66.776	51.523	58.166
XXI	0	0	0.000	—	—	—
XXII	0	0	0.000	—	—	—
Total:	86, 554	28, 852	Average:	73.498	69.614	71.031

Some of the previous research on coding death certificates has focused on deaths related to cancer [5]. When considering the 20 most common ICD cancer blocks in our test split, Model (iii) achieves a macro-averaged F1-score of 90.090%.

Although the results on Table 2 fail to show that neural attention mechanisms lead to an increased performance, these methods can offer model interpretability, by allowing us to see which parts of the input (i.e., which fields and which words) are attended to when making predictions. In Fig. 3, we illustrate the attention weights calculated as shown in Eq. 6, for the contents of two death certificates. The certificate in Fig. 3a was correctly assigned to code C719 (i.e., malignant neoplasm of brain, unspecified) with a confidence of 95.21%, and the figure shows the words *glioblastoma multiforme* having a significant impact. The certificate in Fig. 3b was assigned to code J40 (i.e., bronchitis, not specified as acute or chronic) with a confidence of 92.39%. In this example, the words *insuficiência*

Table 4. Results for blocks and full codes within ICD Chapters II and IX.

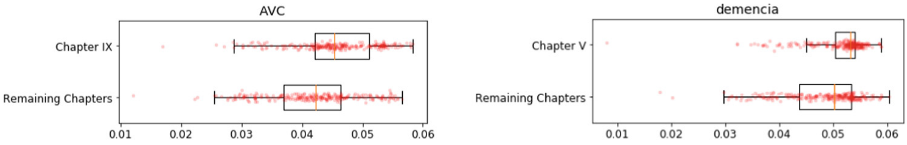
	ICD-10 level	Accuracy	Macro-averages		
			Precision	Recall	F1-score
Chapter II	Block	89.673	31.692	27.226	28.619
	Full code	85.216	26.133	23.877	23.960
Chapter IX	Block	76.859	13.596	11.118	11.900
	Full code	73.683	13.162	10.786	11.221



(a)

(b)

Fig. 3. Examples of the attention weights given at the field and word levels.



(a)

(b)

Fig. 4. Distribution of attention weights given to tokens *AVC* and *demencia*.

cardíaca descompensada in the first field have much less impact than the word *traqueobronquite* on the second field. Figure 4 instead shows the distribution of the attention weights for two particular word tokens, contrasting 250 death certificates from an ICD chapter related to the tokens, against 250 certificates from the remaining chapters. The token *AVC* (i.e., abbreviation of *acidente vascular cerebral*) is often used to denote a stroke, and the attention weights in Chapter IX (i.e., diseases of the circulatory system) are generally higher, as shown in Fig. 4a. Figure 4b shows a similar example, with the token *demencia* and considering Chapter V (i.e., mental and behavioural disorders).

5 Conclusions and Future Work

In this paper, we proposed a deep learning method for coding the free-text descriptions of the cause(s) of death, included in death certificates obtained from the Portuguese Ministry of Health's Directorate-General of Health, according to ICD-10. Results show that although IDC coding is a difficult task, due to the large number of classes that are sparsely used, we can still obtain a high accuracy, particularly in the cases of the more common causes of death. We argue that our approach can indeed contribute to a faster processing of death certificates, or it can help in the task of manual coding. The attention mechanisms used in our complete model also offer the opportunity to interpret and visualize the classification results, as we can check for each input where the model places more attention and how that impacts the prediction. This last aspect is particularly important for applications involving a human in-the-loop (i.e., a health technician with experience in ICD-10 coding), validating the results of the classifier.

Despite the interesting results, there are still many opportunities for future work. For instance, although previous studies have advanced methods for ICD coding of death certificates, their results are not directly comparable to ours, given the different languages and different formulations of the task – in some cases, the input was a single text, and the prediction tasks also differed in the number of classes or in the fact that multiple labels could be given as output. For future work, we would like to experiment with an adapted version of our neural architecture, over the French dataset from the CLEF eHealth task [8].

Noting that the inclusion two different model outputs (i.e., the ICD-10 full code and the ICD-10 block, for the main cause of death) helped to increase accuracy, for future work we would like to further pursue related ideas by considering multiple outputs corresponding to auxiliary causes of death (i.e., in Fig. 1, one can see that in SICO the input strings *(a)* to *(d)* are individually assigned to ICD-10 codes), also leveraging techniques for exploring class co-occurrences [14]. Given the highly skewed class distribution, we also plan to explore batch training procedures that, taking inspiration on the SMOTE method [15], oversample the minority classes and introduce minor perturbations on these training instances.

Finally, we have that the current model is only exploring six small strings as input, although in some circumstances (e.g., accidents, suicide, or homicide) we could also use the supplemental autopsy report. Currently ongoing efforts, also taking inspiration on previous work on text classification [16], are exploring the extension of the deep neural network introduced in Sect. 3 with different parts for encoding the full-text contents of autopsy reports, or the full-text contents of supplemental clinical information bulletins, when these are available.

Acknowledgements. This work had support from Fundação para a Ciência e Tecnologia (FCT), through the INESC-ID multi-annual funding from the PIDDAC program (UID/CEC/50021/2013).

References

1. Marques, C., Maia, C., Martins, H., Pinto, C.S., Anderson, R.N., Borralho, M.D.C.: Improving the mortality information system in portugal. Eurohealth **22**(2), 48–51 (2016)

2. Dalianis, H.: Clinical text retrieval - an overview of basic building blocks and applications. In: Paltoglou, G., Loizides, F., Hansen, P. (eds.) Professional Search in the Modern World. LNCS, vol. 8830, pp. 147–165. Springer, Cham (2014). doi:10.1007/978-3-319-12511-4_8
3. Zweigenbaum, P., Lavergne, T.: Hybrid methods for ICD-10 coding of death certificates. In: Proceedings of International Workshop on Health Text Mining and Information Analysis (2016)
4. Mujtaba, G., Shuib, L., Raj, R.G., Rajandram, R., Shaikh, K., Al-Garadi, M.A.: Automatic ICD-10 multi-class classification of cause of death from plain-text autopsy reports through expert-driven feature selection. PLoS ONE **12**(2), e0170242 (2017)
5. Koopman, B., Zuccon, G., Nguyen, A., Bergheim, A., Grayson, N.: Automatic ICD-10 classification of cancers from free-text death certificates. Int. J. Med. Inform. **84**(11), 956–965 (2015)
6. Koopman, B., Karimi, S., Nguyen, A., McGuire, R., Muscatello, D., Kemp, M., Truran, D., Zhang, M., Thackway, S.: Automatic classification of diseases from free-text death certificates for real-time surveillance. BMC Med. Inform. Decis. Making **15**(1), 53 (2015)
7. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth evaluation lab 2016. In: Fuhr, N., Quaresma, P., Gonçalves, T., Larsen, B., Balog, K., Macdonald, C., Cappellato, L., Ferro, N. (eds.) CLEF 2016. LNCS, vol. 9822, pp. 255–266. Springer, Cham (2016). doi:10.1007/978-3-319-44564-9_24
8. Lavergne, T., Névéol, A., Robert, A., Grouin, C., Rey, G., Zweigenbaum, P.: A dataset for ICD-10 coding of death certificates: creation and usage. In: Proceedings of the Workshop on Building and Evaluating Resources for Biomedical Text Mining (2016)
9. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (2016)
10. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches (2014). arXiv preprint arXiv:1409.1259
11. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Cogn. model. **5**(3), 1 (1988)
12. Kingma, D., Adam, J.B.: A method for stochastic optimization. In: Proceedings of the International Conference for Learning Representations (2015)
13. Goldberg, Y.: A primer on neural network models for natural language processing. J. Artif. Intell. Res. **57**, 345–420 (2016)
14. Kurata, G., Xiang, B., Zhou, B.: Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (2016)
15. Chawla, N.V., Bowyer, K.W., Hall, L.O., Philip, W., Kegelmeyer, S.: Synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
16. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification (2016). arXiv preprint arXiv:1607.01759