Automatic coding of death certificates to ICD-10 terminology

Jitendra Jonnagaddala[1,2,*] and Feiyan Hu[3]

[1] School of Public Health and Community Medicine, UNSW Sydney, Australia
[2] Prince of Wales Clinical School, UNSW Sydney, Australia
z3339253@unsw.edu.au

[3] Insight Centre for Data Analytics, Dublin City University, Ireland
feiyan.hu@dcu.ie

Abstract. In this study, we present methods to automatically assign ICD-10 codes to short plain text description extracted from death certificates in English. We deployed an approach to tackle the task by solely using dictionary lookup, also known as dictionary matching or dictionary projection. The first step is to index manually coded ICD-10 lexicon followed by dictionary matching. Priority rules are applied to retrieve the relevant entity/entities and their corresponding ICD-10 code(s) given free text cause of death description. Because of the dictionary based method that we applied, we were able to evaluate our method even on the training set. The advantages of a dictionary look up method include speed and no need for training data. We present our results of 3 different experimental settings each of which has 2 individual runs. The performance is evaluated by precision, recall and F-measure. We identified several major issues in the corpus contributing to the low performance of our methods. This reiterates the fact that the quality of lexicon plays a significant role on the performance of dictionary lookup based methods.

Keywords: Death certificates coding, Cause of death coding, ICD-10 coding, ICD-10 code assignment, Concept normalization, String matching,

1 Introduction

ICD also known as the International List of Causes of Death, was adopted by the International Statistical Institute in the year 1893[1]. ICD includes the universe of diseases, disorders, injuries and other related health conditions, listed in a comprehensive, hierarchical way to facilitate storage, retrieval, analysis and exchange of information. It is one of the widely used international standards to report diseases and health conditions and to identify health trends and statistics globally. Uses of ICD include monitoring of

[*] Corresponding author

the incidence and prevalence of diseases, observing reimbursements and resource allocation trends, and keeping track of safety and quality guidelines. Another important use is to report deaths as well as diseases, injuries, symptoms, reasons for encounter, factors that influence health status, and external causes of disease.

World Health Organization (WHO published the 6th version in 1948 which is known as ICD-6. All member states of the WHO are regulated to use the most current ICD revision to report mortality and morbidity statistics. The ICD has been revised and published in a series of editions to reflect advances in health and medical science over time. The current ICD version is ICD-10, which was initially used in 1990. It covers more than 20,000 codes including diagnoses and procedures, but only a subset of these codes can be causes of death. Although delayed, ICD-11 is being currently drafted and is expected to be released in 2017.

Manually assigning ICD codes to a free text description is expensive and time-consuming due to the vast coverage and size of ICD terminology, thus automated methods are required to assist the manual coders and public health reporting officials[2]. We can consider the process of assigning ICD codes as a classification problem, or entity recognition problem or, entity recognition and normalization problem, depending on the context. This will allow us to leverage various techniques based on machine learning and/or natural language processing. In recent automatic object detection tasks in images, we have seen deep learning based neural networks outperforming human players [3]. It is legitimate to hypothesize that the ICD code assigning task in future could be completely automated.

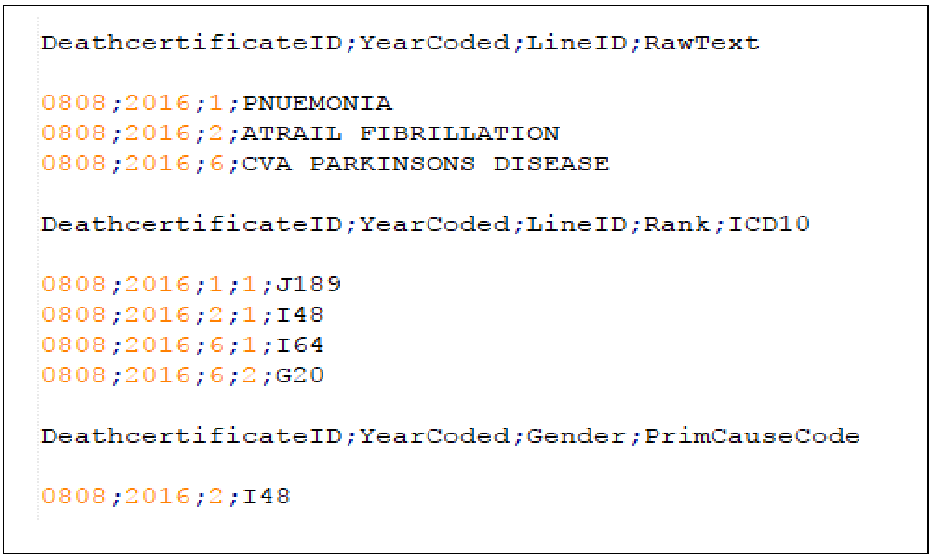
Researchers have been investigating ICD code assignment in different types of medical records such as pathology reports, discharge summaries and death certificates. Recent studies proposed various methods specifically for ICD code assignment in death certificates. Recently, supervised learning methods using Support Vector Machines (SVM) to assign ICD codes has been applied [4-6]. Methods in unsupervised manner are used in few other studies [7, 8]. The methods applied not based on classification models are normally based on dictionary lookup, also known as dictionary matching or projection. Mottin *et al.* used entity relocation and entity normalization to automatically categorize text, compute similarity metric like cosine similarity of features in order to find and rank input text [7]. The feature vector is formed by TF-IDF weighted bag of words. Others claim that the hybrid method of dictionary projection and supervised learning can outperform both dictionary projection and supervised learning [4, 6]. Our method is based on dictionary lookup and priority rules. We applied exact and partial string matching to look up a manually coded ICD-10 dictionary. The result is the corresponding ICD codes of the matching query in the dictionary. The performance of dictionary projection is conditioning on the fact that provided lexicon has good quality. The advantage of such method is that it is easy and cheap to compute on a large scale dataset.

2 Methods

2.1 Corpus

We have used the CDC, distributed as part of the CLEF e-Health 2017 Task 1, for developing our methods[9, 10]. The corpus included censored free-text descriptions of causes of death reported by the clinicians in death certificates. These free-text descriptions were manually coded by the experts using ICD-10 terminology[11]. A manually curated ICD-10 lexicon was provided with the corpus. The methods employed in the construction of this corpus are the same as CépiDc Causes of Death French corpus [12]. The corpus comprised of training and test sets.

A sample (with modified content) ICD-10 coded death certificate from the corpus is shown in Fig.1. In the sample death certificate with ID 0808, there were 3 causes of death entities with ICD-10 codes assigned at line 1, 2 and 6 of the original full death certificate (i.e. “pneumonia”, “atrial fibrillation”, and “CVA parkinsons disease”). There were two ICD-10 codes assigned and ranked manually by the experts for the cause of death statement – “CVA PARKINSONS DISEASE”. The primary cause of death is coded as I48, which stands for “Atrial fibrillation and flutter” in ICD-10 standard terminology. In this study, we only focused on coding all the entities observed in the death certificate. The identification of primary cause of death is beyond the scope of this study. It is also important to note that the corpus didn’t include the full original contents of the death certificates rather it just included only ‘cause of death’ entities.



```
DeathcertificateID;YearCoded;LineID;RawText

0808;2016;1;PNUEMONIA
0808;2016;2;ATRAIL FIBRILLATION
0808;2016;6;CVA PARKINSONS DISEASE

DeathcertificateID;YearCoded;LineID;Rank;ICD10

0808;2016;1;1;J189
0808;2016;2;1;I48
0808;2016;6;1;I64
0808;2016;6;2;G20

DeathcertificateID;YearCoded;Gender;PrimCauseCode

0808;2016;2;I48
```

Fig. 1. Sample ICD-10 coded death certificate.

2.2 Concept coding using dictionary lookup and priority rules

The proposed methods are based on our previous work, on coding PubMed articles with MeSH terminology [13, 14]. Our methods are mainly based on dictionary lookup and priority rules. String matching is a critical technique for dictionary lookup, which can either be exact or partial matching (i.e. proximity and fuzzy matching). The dictionary lookup approach has various advantages and can provide competitive results when used with the right lexicon [14].

Initially, the ICD-10 lexicon and the input free text descriptions in the corpus are subjected to a few pre-processing steps. The pre-processing included tokenization, lemmatization and stop words removal using the Apache Lucene* library. This is followed by the expansion of abbreviations identified in the free text descriptions based on the abbreviations lexicon. This lexicon was developed by the authors in a previous study[14]. Finally, the dictionary matching is performed between the ICD-10 lexicon and the free text descriptions. To identify the right code, we implemented several priority rules. Highest priority is given to the code with an exact match, followed by partial phrase match and partial token match. In many situations, more than one code is identified by each rule, thus we employed another rule to consider only the top code retrieved which had the highest score. The highest score should be greater than 0.5. Similar methods have been employed in a previous study where dictionary look up was used in conjunction with priority rules [7]. However, in our study the priority rules are not just limited to exact match but also cover phrase and term matches.

2.3 Experimental Setup

The training set from the corpus was used to perform initial experiments. The methods discussed in the above section were later evaluated on the test set. Three different experiments (Exp1, Exp2, Exp3), each with two runs (Run1, Run2) were performed on the test set. In each experiment, Run1 refers to the setup where Okapi BM25 scoring was used and TF-IDF scoring for Run2 to rank the retrieved ICD-10 codes[15].

Exp1 considered only ICD-10 codes retrieved which met the priority rule conditions and had the highest-ranking score. No lemmatization and stop word removal steps were employed. Exp2 considered only ICD-10 codes retrieved which met the priority rule conditions and had highest-ranking score. However, lemmatization and stop word removal steps were employed. Exp3 was very similar to Exp2 except with the addition of abbreviation expansion component. We developed a separate lexicon which included abbreviations with their full forms using MEDIC vocabulary [16]. Exp1 and Exp2 were performed on the test set solely based on our initial experiments on the train set. In other words, we didn't access the ground truth of the test set while performing these experiments. Exp3 was performed after performing error analysis on the predicted ICD-10 codes from previous experiments by accessing the ground truth of the test set.

* http://lucene.apache.org/core/

2.4 Evaluation metrics

The performance of the proposed methods was assessed using the standard metrics precision (P), recall (R) and F-measure (F) by identifying the true positives (TP), false positives (FP) and false negatives (FN).

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F = \frac{(2 \times P \times R)}{P + R} \tag{3}$$

The metrics by default consider all the ICD-10 codes irrespective of their type or group in the terminology. However, the metrics were also used to evaluate performance on violent deaths type (codes from V01 to Y98) of ICD-10 codes. The intuition behind evaluating the performance of this type was specifically that public health professionals in general are keen to identify, analyze and intervene in these avoidable deaths. Only Exp2 runs were evaluated for violent deaths type.

3 Results

The above proposed automatic methods were applied to all the death certificates in the dataset. The distribution of the training and test sets of the corpus is summarized in Table 1. We noticed that the performance on the test set is lower than initial experiments performed on the training set.

Table 1. Distribution of train and test sets.

	Training set	Test set
No. of death certificates	13,330	6,665
No. of entities in all death certificates	40,351	18,444
No. of ICD-10 codes (excluding those without codes)	39,332	18,928
No. of entities without ICD-10 codes	2252	119
No. of unique ICD-10 codes	1255	900
No. of tokens* in all death certificates	96,177	45,354
Average token count per death certificate	7.22	6.81

* Tokens are calculated using NLTK tokenizer http://www.nltk.org/

The results of our experiments described in the previous section on the test set are presented in the Table 2. In Exp2 and Exp3 the BM25 scoring based Run1 outperformed TF-IDF scoring based Run2. The performance results specifically for violent deaths type for Exp2 runs were as follows, Exp2-Run1 achieved 0.1684(P), 0.2619(R) and 0.205(F), while Exp2-Run2 achieved 0.043(P), 0.3095(R) and 0.0755(F).

Table 2. Performance results on the test set

Setup	Evaluation metrics					
	TP	FP	FN	P	R	F
Exp1-Run1	8112	19137	10666	0.2977	0.4320	0.3525
Exp1-Run2	7915	17870	10863	0.3070	0.4215	0.3552
Exp2-Run1	6607	9891	12171	0.4005	0.3518	0.3746
Exp2-Run2	6156	10441	12622	0.3709	0.3278	0.3480
Exp3-Run1	7090	9604	11688	0.4247	0.3776	0.3998
Exp3-Run2	6605	10081	12173	0.3958	0.3517	0.3725

4 Discussion

Our results demonstrate that the performance of dictionary lookup based approach for ICD-10 code assignment in death certificates is inferior to supervised and/or hybrid based methods [5, 6]. To identify the possible reasons for large number of FN and FP, a thorough error analysis was manually performed on a subset of predicted ICD-10 codes based on the Exp2 setup. Many issues were noticed ranging from quality of the lexicon supplied in the corpus to short comings in our experimental setup. One of the short-comings we addressed was addition of abbreviation expansion as part of the Exp3. We identified that the testing set and training set included various abbreviated ‘cause of death’ entities which were not addressed in Exp1 and Exp2. HTN, CAD, COPD, CHF, CAR and CVA were some of the frequently abbreviated entities appearing in the death certificates. Our custom abbreviations lexicon had around 350 entries and it increased our F score from 0.3746 to 0.3998.

One of the key reason for our low performance was quality of the ICD-10 lexicon supplied. We observed many issues including inconsistent formatting errors and incomplete coverage of ICD-10 codes in the lexicon. For example, we noticed that there were over 100 instances where the ICD-10 codes manually coded by the experts are not part of the ICD-10 lexicon. W19, W75 and B334 were few such examples observed in the corpus. There were also several issues noticed with coding performed by the experts. There were inconsistencies in the lexicon and codes identified manually by the experts. One such example is that there are instances where experts coded few entities to J101 but in the lexicon the correct corresponding code is J1010. Another similar type of issue is the ‘cause of death’ entities in a death certificate don’t match to the expert coded version. For example, consider the death certificate with ID 00004. There is only one entity (STROKE) according to the file which doesn't include ICD-10 codes but in the expert coded version there were two (I64 and F179) ICD-10 codes.

Inconsistencies in the representation of multiple entities observed in the same line of the death certificate were also frequently observed throughout the corpus. "CVA PARKINSONS DISEASE" is such example where Cerebrovascular accident (CVA) and PARKINSONS DISEASE are not clearly separated. "H/O CAD AND ELEVATED B/P", "Respiratory Distress/arrest", "HEMORRHAGE S/P AORTOBIFEMORAL BYPASS", "CHF - DIASTOLIC" and "H/O CAD AND ELEVATED B/P" are similar such examples where entities are separated inconsistently with no standard guidelines or notation. There were at least over 2000 instances of such inconsistencies in both train and test sets. We strongly believe that by enhancing the current ICD-10 lexicon, we can improve the dictionary lookup based performance further. One enhancement worth exploring in future is to incorporate synonyms and, spelling variations and corrections (Example: PNUEMONIA => PNEUMONIA; ATRAIL FIBRILLATION => ATRIAL FIBRILLATION) into ICD-10 lexicon used in addition to addressing the issues discussed earlier.

5 Conclusion

In conclusion, we have described our methods to automatically code death certificates to ICD-10 terminology. Our dictionary-lookup based methods are simple, effective and no training phase is required. However, the performance of these methods is not as good as machine learning based topic modeling or learning to rank or hybrid methods. The performance of dictionary lookup heavily relies on the quality of the lexicon used. In addition, to a high-quality lexicon, enhancements such as synonym and spelling variations need to be incorporated into dictionary lookup approach for better performance. In future, we would like to improve our results by employing learning to rank algorithms in conjunction with improved dictionary lookup approach.

Acknowledgements

This study was conducted as part of the electronic Practice Based Research Network (ePBRN) and Translational Cancer research network (TCRN) research programs. ePBRN is funded in part by the School of Public Health & Community Medicine, Ingham Institute for Applied Medical Research, UNSW Medicine and South West Sydney Local Health District. TCRN is funded by Cancer Institute of New South Wales and Prince of Wales Clinical School, UNSW Medicine. We would like to thank the organizers of CLEF eHealth 2017 Task 1 for providing us, with the ICD10 coded text content from death certificates. The content of this publication is solely the responsibility of the authors and does not necessarily reflect the official views of the funding bodies.

References

1. WHO. [cited 2017 June]; Available from: http://www.who.int/classifications/icd/en/HistoryOfICD.pdf.
2. Jonnagaddala, J., et al., *Mining Electronic Health Records to Guide and Support Clinical Decision Support Systems*, in *Improving Health Management through Clinical Decision Support Systems*. 2016, IGI Global. p. 252-269.
3. He, K., et al. *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*. in *Proceedings of the IEEE international conference on computer vision*. 2015.
4. Boytcheva, S. *Automatic matching of ICD-10 codes to diagnoses in discharge letters*. in *Proceedings of the Workshop on Biomedical Natural Language Processing*. 2011.
5. Dermouche, M., et al. *ECSTRA-INSERM@ CLEF eHealth2016-task 2: ICD10 code extraction from death certificates*. 2016. CLEF.
6. Zweigenbaum, P. and T. Lavergne, *Hybrid methods for ICD-10 coding of death certificates.* EMNLP 2016, 2016: p. 96.
7. Mottin, L., et al., *BiTeM at CLEF eHealth Evaluation Lab 2016 Task 2: Multilingual Information Extraction.*
8. Zweigenbaum, P. and T. Lavergne. *LIMSI ICD10 coding experiments on CépiDC death certificate statements*. 2016. CLEF.
9. Lorraine Goeuriot, L.K., Hanna Suominen, Aurélie Névéol, Aude Robert, Evangelos Kanoulas, Rene Spijker, João Palotti, and Guido Zuccon. , *CLEF 2017 eHealth Evaluation Lab Overview.* , in *CLEF 2017 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September, 2017*. 2017.
10. Névéol, A.a.A., Robert N. and Cohen, K. Bretonnel and Grouin, Cyril and Lavergne, Thomas and Rey, Grégoire and Robert, Aude and Rondet, Claire and Zweigenbaum, Pierre. , *CLEF eHealth 2017 Multilingual Information Extraction task overview: ICD10 coding of death certificates in English and French.* , in *CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, September, 2017.*
11. WHO, *The ICD-10 Classification of Diseases, Clinical Descriptions and Diagnostic Guidelines.* Geneva: WHO, 1992.
12. Lavergne, T., et al., *A Dataset for ICD-10 Coding of Death Certificates: Creation and Usage.* BioTxtM 2016, 2016: p. 60.
13. Jonnagaddala, J., et al. *Recognition and normalization of disease mentions in PubMed abstracts*. in *Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain, September 9-11, 2015.* 2015.
14. Jonnagaddala, J., et al., *Improving the dictionary lookup approach for disease normalization using enhanced dictionary and query expansion.* Database, 2016. **2016**: p. baw112-baw112.
15. Manning, C.D., P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Vol. 1. 2008: Cambridge university press Cambridge.
16. Davis, A.P., et al., *MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database.* Database, 2012. **2012**: p. bar065.