Artificial Intelligence In Medicine xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

# Extracting cancer mortality statistics from death certificates: A hybrid machine learning and rule-based approach for common and rare cancers

Bevan Koopman[a,*], Guido Zuccon[b], Anthony Nguyen[a], Anton Bergheim[c], Narelle Grayson[c]

[a] The Australian e-Health Research Centre, CSIRO, Brisbane, Australia
[b] Queensland University of Technology, Brisbane, Australia
[c] Cancer Institute NSW, Sydney, Australia

ARTICLE INFO

*Keywords:*
Cancer classification
Death certificates
Machine learning
Natural language processing
Rules
Hybrid

ABSTRACT

*Objective:* Death certificates are an invaluable source of cancer mortality statistics. However, this value can only be realised if accurate, quantitative data can be extracted from certificates—an aim hampered by both the volume and variable quality of certificates written in natural language. This paper proposes an automatic classification system for identifying all cancer related causes of death from death certificates.

*Methods:* Detailed features, including terms, *n*-grams and SNOMED CT concepts were extracted from a collection of 447,336 death certificates. The features were used as input to two different classification sub-systems: a machine learning sub-system using Support Vector Machines (SVMs) and a rule-based sub-system. A fusion sub-system then combines the results from SVMs and rules into a single final classification. A held-out test set was used to evaluate the effectiveness of the classifiers according to precision, recall and *F*-measure.

*Results:* The system was highly effective at determining the type of cancers for both common cancers (*F*-measure of 0.85) and rare cancers (*F*-measure of 0.7). In general, rules performed superior to SVMs; however, the fusion method that combined the two was the most effective.

*Conclusion:* The system proposed in this study provides automatic identification and characterisation of cancers from large collections of free-text death certificates. This allows organisations such as Cancer Registries to monitor and report on cancer mortality in a timely and accurate manner. In addition, the methods and findings are generally applicable beyond cancer classification and to other sources of medical text besides death certificates.

## 1. Introduction

Cancer notification and reporting remains a critical activity for Cancer Registries who are charged with providing an accurate picture of the impact of cancer, the effect of cancer treatments and to direct research efforts for cancer control. A critical source of cancer information comes in the form of free-text death certificates [1]. Death certificates provide population-based cancer mortality statistics that in turn provide a measure of the effectiveness of healthcare systems and guide cancer control strategies [2].

However, Cancer Registries receive an overwhelming number of death certificates (about 44,700 certificates annually for the Cancer Institute NSW[1]); only a portion of these contain cancer (approx. 30% [3]). Manual identification of cancers from this volume of certificates is resource intensive. An effective automated method for cancer classification would allow for up-to-date mortality information used in the monitoring, planning and evaluating the management of cancers that are of high public health importance. Some automated approaches have been developed [4], however, these are typically targeted at specific cancers and do not consider an integrated system that includes all cancers, both common and rare.

In this paper, we propose an integrated system for the automatic classification of all cancers—both common and rare—from free-text death certificates. The system has a number of components: (i) a natural language processing (NLP) pipeline that extracts detailed features (e.g., terms, n-grams, SNOMED CT[2] codes and ICD-O[3] properties) from death

* Corresponding author at: UQ Health Sciences Building, Royal Brisbane Hospital, Herston, Queensland 4029, Australia.
*E-mail addresses:* bevan.koopman@csiro.au (B. Koopman), g.zuccon@qut.edu.au (G. Zuccon), Anthony.Nguyen@csiro.au (A. Nguyen), anton.bergheim@cancerinstitute.org.au (A. Bergheim), Narelle.Grayson@cancerinstitute.org.au (N. Grayson).

[1] Annual average for years 1999–2008, obtained using the dataset from this study.

[2] SNOMED CT is a standardised healthcare terminology including comprehensive coverage of diseases, clinical findings, therapies, procedures and outcomes.

[3] The International Classification of Diseases for Oncology (ICD-O) is a domain-specific extension of the International Statistical Classification of Diseases and Related Health Problems for tumor diseases. This classification is widely used by cancer registries.

https://doi.org/10.1016/j.artmed.2018.04.011
Received 12 June 2017; Received in revised form 26 April 2018; Accepted 30 April 2018

certificates; and (ii) a set of machine learning classifiers that exploit these features to determine the presence of common cancers; (iii) a set of rule-based methods for better handling rare cancers; and (iv) a fusion method to combine the machine learning and rule-based methods into a single system (see Fig. 3 for an architectural overview).

A detailed empirical evaluation on 10 years of coded death certificates shows that the proposed system is effective at determining the type of cancers for both common cancers (*F*-measure of 0.85) and rare cancers (*F*-measure of 0.7). Overall combined *F*-measure effectiveness was 0.84.

Analysis of the results shows that many death certificates received multiple positive cancer classifications from different classifiers (both rule and SVM), whereas a requirement was to determine a single underlying cause of death. The proposed fusion method overcomes this by applying a number of different strategies to rank multiple classifications and determine the most likely, single classification.

The findings of this study helps guide the development of automated methods for multi-class text classification tasks beyond cancer classification and could be applied to other data sources besides death certificates.

## 2. Task description—identifying cancer from death certificates

The use case or task proposed in this study is to identify whether a specific cancer (according to the ICD-10 classification system) was the underlying cause of death from a free-text death certificate. It is important that this works for all cancers, both common and rare, as these can have differing requirements. For common cancers that have a high impact on society, an automated system allows for accurate monitoring to understand and direct treatment efforts. For rare cancers, an automated system provides a means to find rare yet important pieces of information that may help better understand and treat such cancers.

Before detailing in the next sections how this can be achieved with an automated classification system, this section provides an understanding of the particular characteristics of death certificates and the data collection methods used in this study; this helps to understand the design of the automated classification system.

### 2.1. Death certificate format

Death certificates are authored according to a specific procedure [5] and therefore this affects how any automated classification is both developed and evaluated. Fig. 1 provides a sample death certificate. Section (I) contains the main causes of death with the first entry, (A), being the "Disease or condition directly leading to death". The ordering of section (I) should be interpreted as (A) "due to or as a consequence of" (B) "due to..." (C), with the last entry, (D), often (but not always) listed as the *underlying* cause of death. Section (II) contains "Other significant conditions contributing to the death, but not related to the disease condition causing it". For the purpose of this study, this certificate should be classified as of type C34 (*Malignant neoplasm of bronchus and lung*).

| | |
|---|---|
| (I) | A) CARDIORESPIRATORY FAILURE |
| | B) DECOMPENSATED CHRONIC LIVER FAILURE |
| | C) LIVER METATSASIS |
| | D) METASTATIC SMALL CELL LUNG CANCER |
| (II) | EMPHYSEMA |

**Fig. 1.** Sample death certificate. The certificates conforms to a format recommended by the World Heath Organisation, where section (I) contains the causes directly leading to death and (II) contains other contributing conditions.

**Table 1**
Dataset of death certificates; separated into training and test sets based on the year the death certificate was issued.

| | Training set | Testing set |
|---|---|---|
| Years | 1999–2006 | 2007–2008 |
| Num. certificates | 355,165 | 92,171 |
| % cancer | 29.0% | 29.9% |

### 2.2. Collection of death certificates

The Cancer Institute NSW supplied free-text, de-identified death certificates for the years 1999–2008 (inclusive).[4] The certificates were divided into separate training and testing sets so that automatic methods could be developed using certificates from the training set and subsequently evaluated on certificates from the unseen test set. The train/test split was based on the year the certificate was issued, with details provided in Table 1. The split of training and test sets by date was deliberately done because this reflects the realistic setting in which the system would be used in a Cancer Registry. In such a real-world setting, a classifier could only be trained on retrospective data from previous years and then used to classify data from the current year; thus we replicate this situation in our experimental methodology.

### 2.3. Ground truth

A single underlying cause of death (in the form of ICD-10 code [5]) for each certificate was assigned by the Australian Bureau of Statistics (the organisation responsible for maintaining cause-of-death statistic in Australia). These ICD-10 codes constitute the ground truth against which the automated classification method is evaluated. All ICD-10 codes were truncated at the three characters level; for example, the code C34.1 (*Malignant neoplasm: Upper lobe, bronchus or lung*) was converted to simply C34 (*Malignant neoplasm of bronchus and lung*).

Cancer deaths were identified as those certificates assigned any ICD-10 code from ICD-10 Chapter II (*Neoplasms*) [6], including in situ and benign cancers (i.e., all codes in the range"C00" to "D49"). The frequency distribution according to the type of cancer is shown in Fig. 2. The figure shows that a small subset of cancer types make up the vast majority of cancer-caused deaths: the top 20 most prevalent cancers constitute approximately 85% of all cancer deaths. It also shows that there are a large number of rare cancers. While previous work has focused on either the top 20 common cancers [7–9,4,10], or a specific rare cancer [11,4], in this work we aim to investigate a general solution that handles both common and rare cases.

## 3. Related work

Cancer Registries are increasingly turning to automated methods to extract cancer related statistics from increasing volumes of the cancer related data they receive. For example, the Danish Cancer Registry introduced electronic reporting and integration with the patient administrative system [12]; in Australia, the utility of automatically performing cancer notifications and synoptic reporting from pathology and cytology reports have shown to be promising [13]. These case studies show there is both a need and viable use case for automated classification of cancers from cancer registry data.

There have been a number of text mining applications specifically focusing on extracting cancer related information (Spasic et al. [4] provides a comprehensive review of these.) There are two main automated approaches: rule-based and machine learning based. We review

[4] The data was provided with approval from the NSW Registry of Births Deaths and Marriages under NSW Population & Health Services Research Ethics Committee application HREC/11/CIPHS/60|.