Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com

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 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

* Corresponding author at: UQ Health Sciences Building, Royal Brisbane Hospital, Herston, Queensland 4029, Australia.
Email 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)

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
Available online xxx

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.

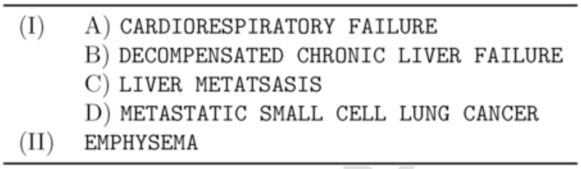
2. Task description—identifying cancer from death certificates

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.

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.

[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.

[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|.

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%

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 the advantages and disadvantages of each below, highlighting the case for a hybrid approach, investigated in this work, that leverages the benefits of both.

Historically, there has been an emphasis towards the use of rule-based techniques [4], i.e., the use of pattern matching and dictionary lookup for cancer-related entity extraction. A number of rule-based approaches make use of natural language processing and pattern matching aided by a medical domain knowledge resource, either the UMLS Metathesaurus [14,15] or some other resource [9]. Rules are typically developed manually, working with a domain expert and via manual review of the textual data being classified.

Machine learning approaches have proven effective in a multitude of different text classification tasks [16], including on death certificates [17]. Specific to cancer classification from death certificates, Butt et al. [18] developed a binary (i.e., cancer or no-cancer) classifier for free-text death certificates. They found that a Support Vector Machine classifier, trained on free-text death certificates with human ground truth data, proved the most effective against a rule-based approach and a number of other classification models. This was limited in that it only identified the cause of death as cancer and did not distinguish between different types of cancers. Furthermore, validation was done on a small 5000 death certificate collection.

This limitation was overcome in a further study [10] by developing classifiers (SVMs) for individual cancer types and empirically evaluating these on a large collection of death certificates. The proposed system had two components: a natural language processing pipeline that extracts features (both term and concept-based) from death certificates; and a series of supervised Support Vector Machines, that utilise the extracted features for classification. The system was effective in classifying common cancers (the same common set as those used in this study, with average *F*-measure of 0.7) but performed poorly on rare cancers (*F*-measure of 0.12)—a major limitation for the use of the system as accurate mortality statistics are needed for both high impact common cancers and for rarer cancers. A further limitation of this approach was that different classifiers could provide a positive classification for a single death certificate (e.g., both breast cancer and lung cancer found to be the cause of death); instead, it is a requirement that a single cause of death be determined (inline with how death certificates are authored).

Both machine learning and rule-based approaches have specific advantages and disadvantages, making a hybrid approach attractive. Rule-based approaches are easy to develop for small amounts of data (e.g., for rare cancers), or when specific cancers are clearly defined and unambiguous. They do not require large amounts of ground truth training data and can be computationally very efficient. However, some limiting factors for rule-based techniques are: the effort required to develop manual rules, which have to be defined on a case-by-case basis for each cancer type; and brittleness of rule-based approaches to the idiosyncrasies of the clinical sublanguage such as non-standard abbreviations as well as a high degree of spelling and grammatical errors. In contrast, machine learning approaches do not require manual effort, can "learn" the idiosyncrasies of the clinical sublanguage and have proved empirically more effective on other classification tasks. The limitations of machine learning are that they do require sufficient training data (a problem for rare cancers) and the feature extraction and model training processes can be computationally expensive.

With each method offering different advantages and disadvantages, this paper investigates a hybrid approach that fused the results from both machine learning and rule-based classifiers. The contributions of this paper are: (i) the development of a symbolic, rule-based approach that utilises both term and concept representations of the death certificate for cancer classification; (ii) a hybrid approach that combines rules and SVMs to leverage the benefits of both approaches; and (iii) the investigation of different "fusion" methods to combine the results (and scores) of multiple different classifiers.

4. Proposed hybrid fusion method

The proposed method is made up of four numbered components, illustrated in Fig. 3: (1) the feature extraction method takes a free-text death certificate and extracts both term and concept (SNOMED CT and ICD-O) features; (2) the machine learning classifiers (SVMs); (3) the set of rules; and (4) the fusion method that combines the classifier scores from the SVMs and rules.

4.1. Feature extraction methods

The feature extraction process was performed using Medtex, a clinical natural language processing system [19]. A variety of different features are used for both the rules and to train a classification model. Features fall into two different categories: (i) basic *term-based* features taken directly from the text of the death certificate; and (ii) *concept-based* features, derived from the original terms, where concepts belong to the medical terminologies SNOMED CT and ICD-O. Table 2 describes the different types of features extracted, belonging to these two categories. For each feature type, a description is provided and an example of the features that are consequently derived given the fragment of a death certificate. While term-based features are commonly used in text classification tasks, the use of SNOMED CT and ICD-O features are more unique and have proven effective [10].

Once all features were extracted, death certificates were transformed from their original terms to feature vectors; for example, each word (TokenStem) or SNOMED CT concept represents a single feature dimension in the vector, with features grouped into high level feature types (TokenStem or SCTConceptId). The actual values in the vector are a binary indication of the presence of the feature in the particular death certificate. Further details on feature selection and their impact on effectiveness can be found in [10].

4.2. Rule-based methods

The rule-based approach used only the SNOMED CT codes extracted from a death certificate as part of the feature extraction method; these were then mapped to ICD-10 codes according to the following procedure:

- SNOMED CT codes were mapped to ICD-O codes. ICD-O is a domain-specific extension of the ICD for tumour diseases [6]. The SNOMED CT to ICD-O maps are provided as part of the SNOMED CT distribution.
- The resulting ICD-O codes were then mapped to ICD-10 codes. This was done via an existing ICD-O to ICD-10 Conversion Program provided by the National Institute of Health but customised to include only mappings to malignant neoplasms (i.e., ICD-10 codes having a

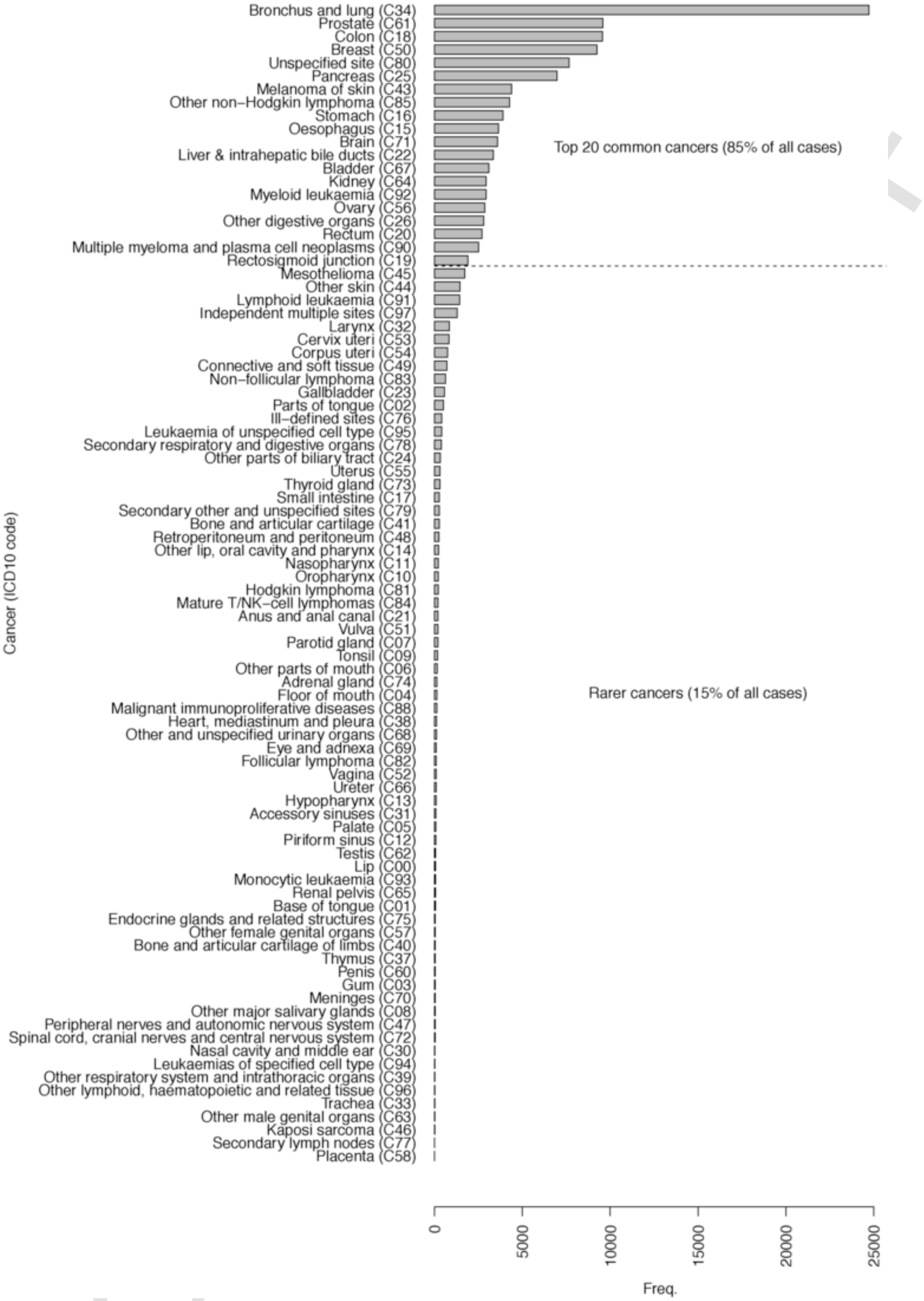


Fig. 2. Number of death certificates with the cause of death classified as cancer by cancer type. Taken from the ground truth for the full set of death certificates (1999–2008).

prefix ‘C’).[5] The mapping table works by defining a mapping {*morphology*, *site*}→ ICD-10 code, i.e., a given cancer morphology and primary site equals to a single cancer ICD-10 code.

- Note that a single death certificate will typically have multiple SNOMED CT concepts extracted from it. This results in multiple ICD-O mapping and, therefore, multiple candidate ICD-10 cancer classifications for a single certificate. However, a single underlying cancer-related cause of death is required. As a result, the candidate ICD-10 codes had to be scored and ranked so that they could be later pruned as part of the fusion method.

4.2.1. Candidate scoring

For each death certificate d, we denote the list of candidate codes as C_d. The score for an individual $c \in C_d$ is made up of three individual scores, interpolated as:

$$\begin{aligned} \text{rule\_score}(c) = {} & \lambda_p \text{prev\_score}(c) + \lambda_s \text{section\_score}(c) \\ & + \lambda_m \text{morph\_site\_score}(c), \end{aligned} \tag{1}$$

[5] ICD-O-3 to ICD-9-CM, ICD-10 (Cause of Death) and ICD-10-CM ICD Conversion Programs http://seer.cancer.gov/tools/conversion.

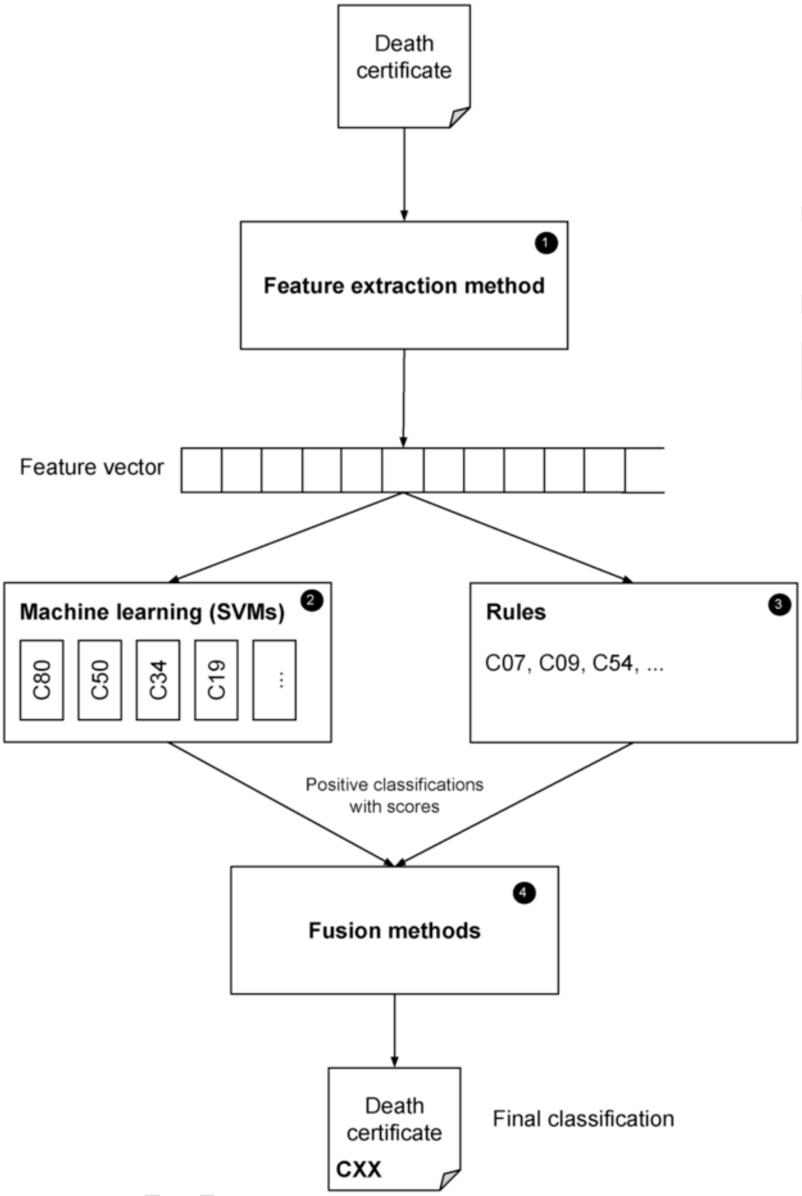


Fig. 3. Architecture overview showing the main components of the proposed cancer classification system. The feature extraction process extracts detailed features, used in turn by both rules and SVMs. Multiple positive classifications for a single death certificates are scored and a single underlying cause of death is determined by the fusion method.

where $\lambda_p + \lambda_s + \lambda_m = 1.0$ and are used to control the relative impact of each source of evidence. These parameters were tuned to the values that maximised *F*-measure on the training set and then applied to the test set.

The *prev _ s core*(*c*) component represents the prevalence of the ICD-10 code *c*:

$$\text{prev\_score}(c) = \frac{N_c}{N}, \tag{2}$$

where N_c is the number of death certificates assigned the ICD-10 code *c* in the training set and *N* is the total number of death certificates in the training set. This effectively represents historical prevalence of cancer *c*.

The *section _ s core*(*c*) component accounts for which section of the death certificate contained the evidence that fired the rule. Death certificate sections are in the following, decreasing order of precedence: I-D, I-C, I-B, I-A, II. To calculate the *section*(*c*) component we first determine the sequence of sections S_d relating to C_d such that each $s_i \in S_d$ represents the section for $c \in C_d$. The sections S_d are then sorted according to the precedence order specified above and duplicates removed. Finally, the section score is then calculated as:

$$\text{section\_score}(c) = \frac{|S_d|-i}{|S_d|}, \tag{3}$$

where $|S_d|$ is the unique number of sections for that death certificate and *i* is the index position of $s_i \in S_d$ with s_0 representing the highest precedence section (e.g., section I-D if that exists in *d*, otherwise the next highest, I-C and so forth).

The *morpth _ site _ score*(*c*) component captures the {*morphology*, *site*}→ ICD-10 code mapping:

$$\text{morph\_site\_score}(c) = \begin{cases} 1.0, & \text{if a single ICD} - 10 \text{ is deduced} \\ \frac{1}{|\text{possible ICD}-10 \text{ codes}|}, & \text{otherwise} \end{cases} \tag{4}$$

Table 2
Types of features—both term and concept-based—extracted from death certificates. (Stemming is a process of removing and replacing word suffixes to arrive at a common root form of the word.)

Feature type	Description	Example certificate extract	Resulting feature values
Term-based			
TokenStem	A token stem, i.e., the stemmed version of a word.	`Acute chronic renal failure`	Acut, chronic, renal, failur.
TokenStem *n*-gram	The *n*-gram formed by *n* adjacent token stems.	`CHRONIC RENAL FAILURE`	chronic renal, renal failur.
Concept-based			
ICDOMorph	Standard ICD-O Morphology classification system	`ADENOCARCINOMA OF THE LUNG`	M-81403
ICDOSite	Standard ICD-O body site classification	`LUNG CANCER`	C34.9 (Malignant neoplasm of Bronchus or lung)
ICDOMorphBerg	Major groupings of carcinomas and non-carcinomas (as defined by [20])	`ADENOCARCINOMA OF THE LUNG`	Adenocarcinomas
ICDOSiteGroup	Course grained body site descriptions	`LUNG CANCER`	C30-C39 (Respiratory system and intratoracic organ)
SCTConceptId	SNOMED CT concept identifier (as extracted by the Medtex system)	`chronic renal failure`	90688005

If a death certificate mentions both a morphology and a site then a single ICD-10 code is deduced and $morph\_site\_score(c) = 1.0$. If only a *morphology* is mentioned then more than one ICD-10 code may be possible and the $morph\_site\_score$ captures this ambiguity.

4.3. Machine learning methods

We reproduced the method of [10], namely multiple ICD-10 binary classifiers, one for each type of cancer, were trained to label a death certificate with a particular ICD-10 code. For the implementation of the classifiers we use Support Vector Machines (SVMs). The feature vectors resulting from the feature extraction process were used to train the SVMs. The training set was taken from 1999 to 2006 certificates, while certificates from 2007 to 2008 were held out as an independent test set (as per Table 1).

In addition to the binary classification of the SVMs, a score or likelihood of correctness is required to handle cases when multiple positive classifications occur for a single death certificate. This SVM likelihood can as a function of the distance from the point to be classified and the hyperplane, which separates positive and negative classifications. This likelihood is used in conjunction with the rule-based score to determine the highest ranked classification to assign to a given death certificate.

4.4. Fusion method

Multiple ICD-10 codes can be applied to a single death certificate, either by multiple SVMs producing positive classifications or by the rule-based approach producing mappings to multiple ICD-10 codes; or a combination of the two. However, there is a requirement for a single underlying cause of death for a given certificate. Therefore, a fusion strategy was developed to produce a single cause of death; this is outlined below.

Let C_d be the set of positive classifications (from rule or SVM) for a particular death certificate d. Note that each positive classification is made of a tuple $(c_i, s(c_i))$, where c_i represents the ICD-10 class and $s(c_i)$ represents the score or likelihood of that classification. The fusion strategy was implemented as a fusion function $F(C_d) : C_d \rightarrow P_d$, where P_d represents the final set of *predictions* for d and $P_d \subseteq C_d$. Technically, P_d should contain the single, final class to assign to a certificate d; however, we did experiment with some fusion strategies that allowed more than one class to be assigned to a death certificate (i.e., a ranking of classes) in order to understand the effect of the requirement to choose a single class. A number of different variants of the fusion function, denoted $F_i(C_d)$, were evaluated; these are outlined in Table 3.

Both SVMs and rules produce a score for their respective classifications (rules via the $rule\_score$ and SVMs via the SVM likelihood). These two different scores were normalised to ensure they were comparable. Normalisation was done separately from each method (SVMs or rules) as follows:

1. All the classifier scores on the training set from the method were sorted into an ordered list. The list was divided into 10 buckets, $b_1 \dots b_{10}$ so that each bucket contained a near equal number of classifier scores.
2. For each classifier score within a bucket b_i, we determined whether the classification was correct or not. This resulted in a probability $P(correct|b_i)$ for all the classifiers that produced a score in that bucket. This was done using 10-fold cross validation on the training set.
3. A function $score(c) \rightarrow P(correct|score(c))$ was used to normalise the score according to which bucket $score(c)$ fell within.

Thus two functions were implemented to map the SVM and rule-based scores to a comparable likelihood that could be used in the fusion method.

4.5. Evaluation measures

Three evaluation measures are considered: precision, recall and *F*-measure. Precision (also called positive predictive value) is the fraction of positively classified certificates that were a specific cancer[6], while recall is the fraction of actual specific cancer certificates that are positively classified.[7] For Cancer Registries, both precision and recall are important: a high precision indicates that the system assigns the right ICD-10 code to a certificate, while a high recall indicates the system does not miss certificates (particularly important for rare cancers). To provide a single, overall evaluation measure, precision and recall are combined into a third evaluation measure, *F*-measure.[8]

For analysis and interpretation of the ICD-10 classification results, results were divided into two sets, constituting *common* and rare *can-*

[6] Precision = True Positives/(True Positives + False Positives).

[7] Recall = True Positives/(True Positives + False Negatives).

[8] *F*-measure = 2 * (precision * recall)/(precision + recall).

Table 3
Different fusion strategy methods for dealing with multiple positive classification for a single death certificate.

$F_{all}(C_d)$	Assign all the positive classifications to the certificate. This strategy is used as a baseline where no fusion is performed.
$F_{max}(C_d)$	Assign the positive classification(s) with highest likelihood.
$F_{maxOne}(C_d)$	Assign the single positive classification with highest likelihood. (If there are more than one then take the last class that is encountered only.)
$F_{th}(C_d, \tau)$	Assign the positive classifications with a likelihood greater than a supplied decision threshold parameter, τ. If there are no classes above the decision threshold then default back to strategy $F_{max}(C_d)$. (τ tuned on the training set and evaluated on the test set.)
$F_{his}(C_d)$	For each positive classification, recalculate the likelihood $P(c) = P(c) * P_h(c)$ where $P_h(c)$ represents the likelihood of that cancer by historic frequency data (calculated based on prevalence in the training set). Once new likelihood are calculated apply $F_{max}(C_d)$.
$F_{hisOne}(C_d)$	For each positive classification, recalculate the likelihood $P(c) = P(c) * P_h(c)$ where $P_h(c)$ represents the likelihood of that cancer by historic frequency data (calculated based on prevalence in the training set). Once new likelihoods are calculated apply $F_{maxOne}(C_d)$.
$F_{hisInt}(C_d, \alpha)$	For each positive classification, recalculate the likelihood by interpolating $P(c) = \alpha * P(c) + (1 - \alpha) * P_h(c)$ where $P_h(c)$ represents the likelihood of that cancer by historic frequency data (calculated based on prevalence in the training set). Once new likelihoods are calculated apply $F_{max}(C_d)$. The α parameter was set to 0.8.
$F_{ran}(C_d)$	Select a random prediction from the list; this is used as a baseline for comparison.

cers. The set of common cancers was derived by: (i) ranking ICD-10 classes in descending order of prevalence (according to the ground truth of the testset); and (ii) selecting the top *k* cancers such that 85% of all cancer deaths were covered. The set of rare cancers was simply those ICD-10 classes not contained in the top *k* common cancers; these constituted the remaining 15% of cancer deaths. (Fig. 2 shows the breakdown of common and rare cancers.)

5. Results & analysis

5.1. Overall classification results

Classification results are shown in Table 4. (Results are shown for fusion strategy F_{maxOne}; the results of other strategies are considered in the next section.)

Common cancers were always easier to classify than rare cancers for all three methods (rules, SVMs and fusion). Recall was higher than precision, indicating that false positives were the main type of error in this classification task. The SVMs method, in particular, exhibited higher recall and lower precision, indicating that false positives were even more prevalent when using SVMs compared to rules.

Generally, rules perform better than SVMs. Although recall was best when using SVMs, precision was far better when using rules. *F*-measure was slightly better for SVMs on common cancers, but *F*-measure was considerably better for rules on rare cancers. On all cancers, rules were best in terms of *F*-measure.

As indicated, in some situations rules were best while in others SVMs were best. This finding was the motivation for the hybrid fusion method. This is confirmed in the empirical findings—the fusion method is more effective than SVMs and rules. The fusion exhibits a 5% improvement in *F*-measure for common cancers, a 1% improvement in *F*-measure for rare cancers, and a 6% improvement in *F*-measure for all cancers, when compared to the best performing SVMs or rules methods.

5.2. Rules parameter sensitivity for sources of evidence

The *rule _ s core* for a rule classifier involved combining three different sources of evidence: the historic prevalence of the cancer, the section of the death certificate where the classification came from and the morphology-site mappings. These sources were interpolated with three parameters controlling the influence of each. Fig. 4 shows the sensitivity for these different parameters and, therefore, the importance of each source of evidence (as evaluated on the test set).

The plots show that all three sources of evidence were needed for effective scoring of rules. Importantly, though, the effectiveness (*F*-measure) was stable for different settings of the interpolation parameters. The best overall setting was $\lambda_p = 0.5$ (prevalence), $\lambda_s = 0.3$ (section) and $\lambda_m = 0.2$ (morph site).

5.3. Fusion analysis

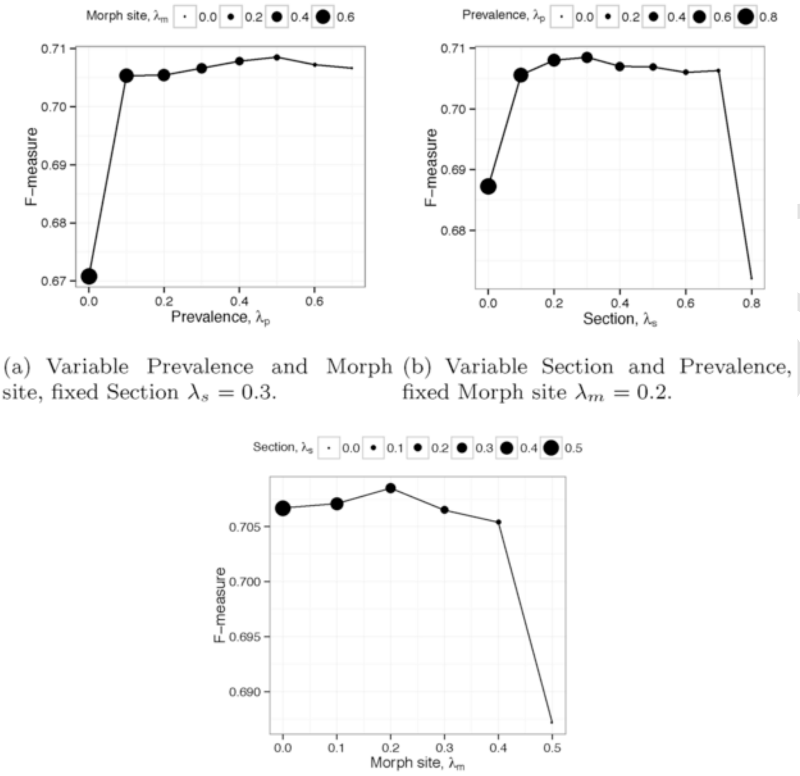
Fusion only occurs when more than one positive classification occurs for a single death certificate. This begs the question of how often this occurred and, on average, how many positive classifications were provided by each method; this is shown in Fig. 5. First, we note that nearly all death certificates received more than one positive classification (i.e., the plot shows very few cases where *x*-axis is 0 and these are solely due to the rules, as expected given that rules had lower recall). As the vast majority of certificates had multiple classifications, the fusion method was applied, and indeed required, in most cases.

The rules generally provided a small number of positive classifications (mean 1.5/certificate), whereas the SVMs generally provided more positive classifications per certificate (mean 7.2/certificate). This large number of classifications also explained the higher number of false positives using SVMs. When SVMs and rules were combined in the fusion method (righthand plot of Fig. 5), the mean number of unique classification per certificate was 7.7; thus, rules provide other possible codes from the SVMs. This shows that fusion was required to resolve the single underlying cause of death for a certificate.

Table 4
Micro and macro-average effectiveness of different classification methods—SVMs, rules and Fusion—for both common and rare cancers. Statistical significance using paired *z*-test: ^{r} indicates significantly better than rules, ^{s} indicates significantly better than SVMs and ^{f} indicates significantly better than Fusion.

	Common			Rare			All		
	Prec	Recall	Fmeas	Prec	Recall	Fmeas	Prec	Recall	Fmeas
Micro-average									
SVMs	0.73	0.91^r	0.81	0.24	$0.84^{r,f}$	0.37	0.61	0.90^r	0.73
Rules	0.78^s	0.82	0.80	0.74^s	0.75	0.75^s	0.77^s	0.82	0.79^s
Fusion	$0.80^{s,r}$	0.90^r	$0.85^{s,r}$	0.74^s	0.79^r	0.76^s	$0.79^{s,r}$	0.89^r	$0.84^{s,r}$
Macro-average									
SVMs	0.55	0.96^r	0.70	0.07	0.90	0.12	0.18	0.91	0.30
Rules	0.80^s	0.79	0.77	0.80^s	0.76	0.76^s	0.79^s	0.77	0.76^s
Fusion	0.82^s	0.87	0.84	0.80^s	0.77	0.77^s	0.80^s	0.81	0.80^s

Results shown for fusion strategy F_{maxOne}.



(a) Variable Prevalence and Morph site, fixed Section $\lambda_s = 0.3$.

(b) Variable Section and Prevalence, fixed Morph site $\lambda_m = 0.2$.

(c) Variable Morph site and Section, fixed Prevalence $\lambda_p = 0.5$.

Fig. 4. Influence of the rule_score parameters controlling the death certificate section, historic prevalence and morphology site mapping. Evaluation on the test set.

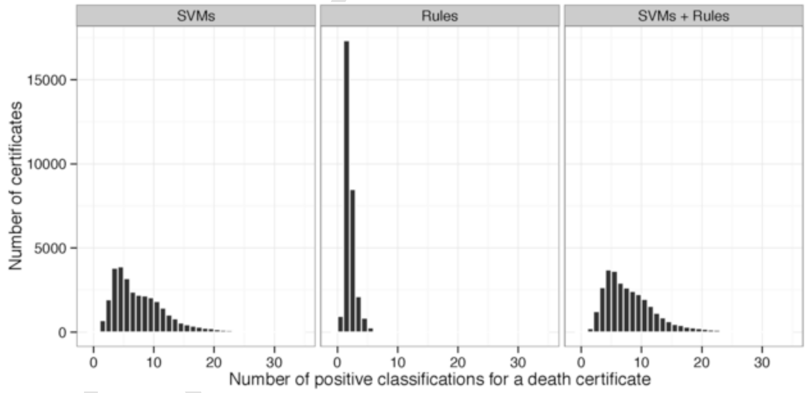


Fig. 5. Distribution of the number of positive classifications provided for a death certificate (using death certificate from the test set). Nearly all death certificates received more than one positive classification. In general, SVMs produced far more positive classifications than rules.

Choosing a single certificate from multiple candidates was done via eight different fusion methods (previously detailed in Table 3). The effectiveness of these different strategies is shown in Fig. 6. There were small variations in effectiveness between the different fusion strategies. The F_{all} method was used as one of the baselines; it kept all the classifications for a certificate. This clearly increased recall as multiple classifications were retained (if any were correct the certificate was deemed correctly classified), but decreased precision as it led to more false positives. The other baseline is F_{ran}, which selected a random class from the list of positive classifications. This was clearly the least effective strategy and thus indicated that a better fusion strategy was required to choose the best class to apply.

The most effective fusion strategy (in terms of *F*-measure) was F_{maxOne}, which chose the single class with the highest score (ties were broken by choosing the last classification encountered). This strategy was slightly better in *F*-measure than F_{max} which did not break ties and instead kept multiple classifications (whereas, a single underlying cause of death is required) with the highest scores. Other strategies included only keeping classifications above a certain threshold (F_{th}). Even tuning the threshold on the training set to the best setting did not prove effective, indicating that the likelihood scores should only be interpreted relative to each other rather than as an absolute measure of confidence. The final strategies that used prevalence statistics (F_{his}, F_{hisOne} and F_{hisInt}) did not prove more effective than the simple F_{maxOne} strategy.

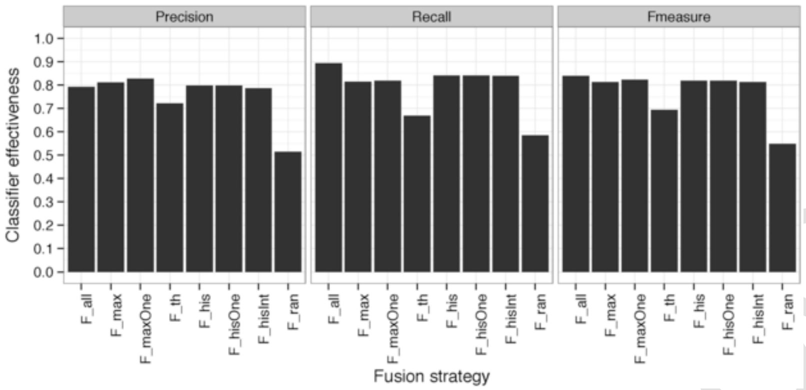


Fig. 6. Effectiveness of different fusion strategies. (Explanation of each fusion strategy provided in Table 3.)

6. Discussion

While rule-based approaches are a common approach for cancer-related text mining [4], the rules proposed here have a number of novel aspects. The rules exploit detailed SNOMED CT and ICD-O features, rather than simple term-based features. This makes them less susceptible to vocabulary mismatch that can occur when relying on terms alone. The rules also rely on three separate sources of evidence (prevalence, section and morphology/site mapping) to provide a confidence score that allows a single classification to be chosen.

On average, each certificate obtained 7.7 positive classifications. The requirement to determine a single cause of death meant that the fusion method was clearly required to handle these multiple classifications. This is an important component that has often been overlooked by other studies [14,15,9,10,17]. In these studies, classifier evaluation is done per-class in isolation (e.g., the performance of individual ICD-10 cancer classifiers [10]). However, in real-world settings all the classifiers will be deployed together within a single system, with a need to handle multiple classifications. In this study, the classifiers are combined and evaluated within a single system.

The fusion method had two major benefits. It overcame the problem of resolving multiple positive classifications and empirically it was statistically significantly better than rules or SVMs alone. Importantly, it was effective on both common and rare cancers.

A hybrid method also provides for some flexibility when new classifiers are required. A choice can be made whether a new classifier should be implemented as a rule or as a SVM. For rare classes, a rule would be most appropriate. Instead, for common classes, with variation in the way the class is expressed in natural language, a SVM may be preferred.

The hybrid approach also allows for flexibility between recall and precision orientated use cases. There are a number of different ways to achieve this using the proposed system. First, the system can make use of either SVMs or rules: for cases where the requirement is to find all cases of a cancer (i.e., high recall) then a SVM may be preferred; for cases where the requirement is to find only the correct cases for one cancer (i.e., high precision) then a rule may be preferred. Second, the $F_{th}(C_d, \tau)$ fusion strategy could be used and the τ decision threshold adjusted to favour either precision or recall tasks. Finally, the ten fold cross validation method employed for classifier score normalisation used *F*-measure, which attributes equal weight to precision and recall. Instead, F_β-score could be used where β controls the relative weight of precision vs. recall. A hybrid system thus allows for some flexibility in extension and deployment of a working system.

In this paper a number of simple fusion strategies were developed. These were mostly heuristic and more advanced approaches could be applied. The positive classifications (and associated scores) can be used as input to a voting model [21], which determines the final classification. Another approach is to use the positive classifications and scores as features to a learning-to-rank [22] model that learns the final ranking.

7. Conclusion

This study provides a system for automatically identifying and characterising cancers—both common and rare—from large collections of free-text death certificates. This allows Cancer Registries to monitor and report on cancer mortality in a timely and accurate manner. 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 codes and cancer specific ICD-O properties) from death 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.

A consequence of having multiple classifiers (both rules or SVMs) deployed is that a single death certificate nearly always receives multiple positive classifications. The system handles this via a number of fusion strategies, with the best strategy being F_{maxOne}, which selects the single classification with the highest score. More advanced fusion strategies that utilise voting or learning methods could be investigated.

The empirical evaluation on 10 years worth of Australian death certificates shows that the system using fusion was effective at determining the type of cancers for both common cancers and rare cancers. The hybrid approach using fusion was better than rules or SVMs alone. The hybrid approach also provides some flexibility in that either rules or SVMs can be preferenced for certain tasks or when adding additional classifiers.

The methods and findings of this study are generally applicable; they can be transferred to other ICD-10 classification tasks beyond cancer classification and to other sources of medical free-text besides death certificates.

Authors' contributions

AB, AN, GZ and NG contributed to the conception of the study. AB and NG constructed the dataset and associated ground truth codes. AN developed the feature extraction methods. BK and GZ developed the machine learning models. BK and AN developed the rules. GZ and BK developed the fusion method. BK performed the empirical evaluation and analysis of results. BK drafted the manuscript. All authors reviewed and approved the final manuscript.

Conflicts of interest

The authors declare that they have no competing interests.

References

[1] R.R. German, A.K. Fink, M. Heron, S.L. Stewart, C.J. Johnson, J.L. Finch, et al., The accuracy of cancer mortality statistics based on death certificates in the United States, Cancer Epidemiol 35 (2) (2011) 126–131.

[2] M. Coleman, D. Forman, H. Bryant, J. Butler, B. Rachet, C. Maringe, et al., Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995–2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data, Lancet 377 (9760) (2011) 127–138.

[3] Cancer Council Australia, Cancer in Australia: facts and figures, 2015 http://www.cancer.org.au/about-cancer/what-is-cancer/facts-and-figures.html, [cited 14.10.15].

[4] I. Spasić, J. Livsey, J.A. Keane, G. Nenadić, Text mining of cancer-related information: review of current status and future directions, Int J Med Inf 83 (9) (2014) 605–623.

[5] World Health Organization, Medical certification of cause of death: instructions for physicians on use of international form of medical certificate of cause of death, 4th ed., World Health Organization, Geneva, 1979.

[6] World Health Organization, International classification of diseases for oncology, (ICD-O-3), 3rd ed., WHO, 2004.

[7] J.C. Denny, N.N. Choma, J.F. Peterson, R.A. Miller, L. Bastarache, M. Li, et al., Natural language processing improves identification of colorectal cancer testing in the electronic medical record, Med Decis Mak 32 (1) (2012) 188–197.

[8] J. Buckley, S. Coopey, J. Sharko, F. Polubriaginof, B. Drohan, A. Belli, et al., The feasibility of using natural language processing to extract clinical information from breast pathology reports, J Pathol Inf 3 (1) (2012) 23.

[9] A.D. Shah, C. Martinez, H. Hemingway, The freetext matching algorithm: a computer program to extract diagnoses and causes of death from unstructured text in electronic health records, BMC Med Inf Decis Mak 12 (1) (2012) 88.

[10] B. Koopman, G. Zuccon, A. Nguyen, A. Bergheim, N. Grayson, Automatic ICD-10 classification of cancers from free-text death certificates, Int J Med Inf 84 (11) (2015) 956–965.

[11] Polpinij J, Miller A, Ontology-based Text Analysis Approach to Retrieve Oncology Documents from PubMed Relevant to Cervical Cancer in Clinical Trials, ICDM Workshop on Advances in Data Mining.

[12] M.L. Gjerstorff, The Danish cancer registry, Scand J Public Health 39 (7 Suppl.) (2011) 42–45.

[13] A.N. Nguyen, J. Moore, J. O'Dwyer, S. Philpot, Assessing the utility of automatic cancer registry notifications data extraction from free-text pathology reports, In: AMIA Annual Symposium Proceedings, Vol. 2015, American Medical Informatics Association, 2015, p. 953.

[14] B. Riedl, N. Than, M. Hogarth, Using the UMLS and simple statistical methods to semantically categorize causes of death on death certificates, In: AMIA Annual Symposium Proceedings, Vol. 2010, American Medical Informatics Association, 2010, p. 677.

[15] K. Davis, C. Staes, J. Duncan, S. Igo, J.C. Facelli, Identification of pneumonia and influenza deaths using the death certificate pipeline, BMC Med Inf Decis Mak 12 (1) (2012) 37.

[16] F. Sebastiani, Machine learning in automated text categorization, ACM Comput Surv 34 (1) (2002) 1–47.

[17] B. Koopman, S. Karimi, A. Nguyen, R. McGuire, D. Muscatello, M. Kemp, et al., Automatic classification of diseases from free-text death certificates for real-time surveillance, BMC Med Inf Decis Mak 15 (1) (2015) 1–10.

[18] L. Butt, G. Zuccon, A. Nguyen, A. Bergheim, N. Grayson, Classification of cancer-related death certificates using machine learning, Aust Med J 6 (5) (2013) 292.

[19] A.N. Nguyen, M.J. Lawley, D.P. Hansen, S. Colquist, A simple pipeline application for identifying and negating SNOMED Clinical Terminology in free text, In: Health Informatics Conference, Canberra, Australia, 2009, pp. 188–193.

[20] J.W. Berg, Morphologic classification of human cancer, in: D. Schottenfeld, J.F. Fraumeni Jr. (Eds.), Cancer epidemiology and prevention, WB Saunders Co, Eastbourne, UK, 1982, pp. 74–89.

[21] C. Macdonald, I. Ounis, Voting techniques for expert search, Knowl Inf Syst 16 (3) (2007) 259–280.

[22] T.-Y. Liu, Learning to rank for information retrieval, Found Trends Inf Retr 3 (3) (2009) 225–331.