Classification of Cancer-related Death Certificates using Machine Learning

Luke Butt[1], Guido Zuccon[1], Anthony Nguyen[1], Anton Bergheim[2], Narelle Grayson[2]

[1]The Australian e-Health Research Centre, Brisbane, Queensland, Australia;

[2]Cancer Institute NSW, Eveleigh, New South Wales, Australia.

RESEARCH

Please cite this paper as: Butt L, Zuccon G, Nguyen A, Bergheim A, Grayson N. Classification of Cancer-related Death Certificates using Machine Learning. AMJ 2013, 6, 5, 292-299. http//dx.doi.org/10.4066/AMJ.2013.1654

Corresponding Author:
Luke Butt
Level 5, UQ Health Science Building 901/16
Royal Brisbane and Women's Hospital
Herston, QLD 4029 Australia
luke.butt@csiro.au

Abstract

Background
Cancer monitoring and prevention relies on the critical aspect of timely notification of cancer cases. However, the abstraction and classification of cancer from the free-text of pathology reports and other relevant documents, such as death certificates, exist as complex and time-consuming activities.

Aims
In this paper, approaches for the automatic detection of notifiable cancer cases as the cause of death from free-text death certificates supplied to Cancer Registries are investigated.

Method
A number of machine learning classifiers were studied. Features were extracted using natural language techniques and the Medtex toolkit. The numerous features encompassed stemmed words, bi-grams, and concepts from the SNOMED CT medical terminology. The baseline consisted of a keyword spotter using keywords extracted from the long description of ICD-10 cancer related codes.

Results
Death certificates with notifiable cancer listed as the cause of death can be effectively identified with the methods studied in this paper. A Support Vector Machine (SVM) classifier achieved best performance with an overall F-measure of 0.9866 when evaluated on a set of 5,000 free-text death certificates using the token stem feature set. The SNOMED CT concept plus token stem feature set reached the lowest variance (0.0032) and false negative rate (0.0297) while achieving an F-measure of 0.9864. The SVM classifier accounts for the first 18 of the top 40 evaluated runs, and entails the most robust classifier with a variance of 0.001141, half the variance of the other classifiers.

Conclusion
The selection of features significantly produced the most influences on the performance of the classifiers, although the type of classifier employed also affects performance. In contrast, the feature weighting schema created a negligible effect on performance. Specifically, it is found that stemmed tokens with or without SNOMED CT concepts create the most effective feature when combined with an SVM classifier.

Word count: 5703
Figures and Tables: 10

Key Words
Death certificates, Cancer Registry, cancer monitoring and reporting, machine learning, natural language processing, SNOMED CT.

What this study adds:

This paper evaluates automatic feature extraction for the automatic machine learning classification of documents from a wide range of tumour streams within a single application. The authors identified salient factors in the composition of machine learning systems and demonstrated the achievement of high classification performances, with the potential to improve workflows for the coding of cancer notifiable free-text death certificates.

Background

Cancer notification and reporting remains an essential and fundamental process for providing an accurate picture of the impact of cancer, the nature and extent of cancer, and to direct research efforts for the cure of cancer. Cancer Registries collect and interpret data from a large number of sources, helping to improve cancer prevention and control,

as well as treatments and survival rates for patients with cancer.

The manual coding of documents, such as pathology reports and death certificates, with respect to notifiable cancers and corresponding synoptic factors (such as primary site, morphology, etc.) exist as a laborious and time consuming process. Cancer Registries strive to provide timely and accurate information on cancer incidence and mortality in the community. These databases receive large quantities of information from a range of sources, such as hospitals, pathology laboratories and Registries of Births, Deaths and Marriages (which issue release death certificates, among others). In addition, there exist cancer mortality cases where only death certificates are available, i.e. no previous pathology or hospital documents reporting cancer were recorded for those individuals. A recent study has reported cases of cancer incidence[1] with death certificates only (DCO) amount to 1.0% and 1.8% in New South Wales and Victoria, respectively (1). Delays in the processing of this data potentially cause underestimation of the incidence of cancer. Computational methods for the automatic abstraction of relevant information possess the ability to enhance a Cancer Registry's workflow, generating time efficiency and costs savings with timely reporting of cancer incidence and mortality information. However, an automatic process is a challenging task, related to the complex nature of the language used in the reports, and the high level of recall and accuracy required.

Previous endeavours attempted to provide automatic cancer coding from free-text pathology reports collected by Cancer Registries. For example, Nguyen et al. employed natural language processing techniques and a rule-based system to automatically extract relevant synoptic factors from electronic pathology reports (2). Likewise, Zuccon et al. demonstrated how these techniques deal with character recognition errors generated by scanning free-text pathology reports stored in paper form (3). Previous research has considered machine learning approaches; for instance, D'Avolio et al. tested approaches based on supervised machine learning (Conditional Random fields and Maximum Entropy) and revealed its effectiveness for the classification of pathology reports in the domains of colorectal, prostate, and lung cancers (4).

Cancer Registries possess access to a number of data sources beyond pathology reports. Death certificates provide one such data supply as a rich source of data that supports cancer surveillance, monitoring and reporting. These certificates contain free-text sections that report the cause of the death of an individual. Figure 1 and 2 provide examples of the free-text content of death certificates where cause of death entails notifiable and non-notifiable cancer types.

[1] The study considered only invasive, primary, malignant neoplasm of the colorectum, lung, breast, or ovary.

Figure 1. De-identified death certificate where cause of death is a notifiable cancer.

(I)A) MAXILLARY TUMOR, 2 YEARS B) PULMONARY OEDEMA, 1 WEEK (II) CEREBROVASCULAR ACCIDENT/DYSPLASIA, 20 YEARS ASTHMA

Figure 2. De-identified death certificate where cause of death is not a notifiable cancer.

I(A) CEREBROVASCULAR ACCIDENT 48 HOURS (B) CEREBRAL ARTERIOSCLEROSIS YEARS (C) HYPERTENSION YEARS II CHRONIC ALCOHOLISM YEARS

Limited research focused on computational methods for automatically classifying death certificates regarding the cause of death. The SuperMICAR system and its related tools provide a semi-automatic coding of the cause of death in death certificates. The system identifies keywords and expressions from the free-text documents that indicate possible causes of death; researchers accomplished this through the use of a standard set of expressions encoded in a predefined vocabulary. Extracted free-text expressions translate to one or more ICD-10 codes which undergo aggregation into a single ICD-10 underlying cause of death through the use of a rule-base. While doctor reported death certificates can be fed directly into the system, Coroner documented deaths require additional pre-processing. A consistent number (between 15 and 20 percent according to a US study) of death certificates cannot be coded through SuperMICAR and related tools, and thus require manual coding (5). Recent research successfully classified death certificates related to pneumonia and influenza using a natural language processing pipeline and rule-based system (6). However, to the best of our knowledge, no previous research investigated fully automatic methods that go beyond keyword spotting of standard cause of death expressions to classifying death certificates, in particular focusing on certificates with cancer as the main cause of death. Furthermore, while Australian Cancer Registries acquire free-text death certificates on a fortnightly basis from the Registry of Births Deaths and Marriages, coded causes of death produced by SuperMICAR (and related products) come from the Australian Bureau of Statistics on a yearly basis. Using computational methods with the ability to tackle the fast identification of death certificates with notifiable cancer as the cause of death potentially produces enhancement of cancer reporting and monitoring capabilities of Cancer Registries.

In this paper, we focus on the problem of automatically identifying death certificates with cancer as the main cause of death. This problem evolves as a binary classification task, i.e. death certificates undergo classification as containing a death cause related to cancer or vice versa as not containing a death cause related to cancer. Several machine learning classifiers are investigated for this task. These include Support Vector Machine, Naive Bayes, decision trees, and boosting algorithms. A state-of-the-art information extraction tool (Medtex) is used to create different set of features to train the classifiers;(7) a number of feature weighting schemas are also considered. Features

encompass stemmed tokens, n-grams, as well as SNOMED CT concept ids and tokens from fully specified names of SNOMED CT concepts, among others. SNOMED CT exist as a medical ontology which formally describes in detail the coverage and knowledge of topics and terminology used in the medical domain (8).

The machine learning classifier approaches being studied underwent testing on 5,000 de-identified death certificates acquired from an Australian Cancer Registry, using 10-fold cross validation to allow for robust training and testing. Our experimental results demonstrate that the choice of weighting schema fails to be critical for achieving high classification effectiveness. Instead, the features chosen to represent the content of death certificates emerge as the primary factor in high classification effectiveness; classifier type appears as a significant secondary factor. Specifically, stemmed tokens arise as the single most important feature set among those extracted. Furthermore, this study found that SNOMED CT features provide consistent increments in classification robustness if used along with stemmed tokens. Although not providing a large increment in classification, the combined use of stemmed tokens bigrams and SNOMED CT concepts provide the lowest range, variance and false negative rate in these experiments.

Next, the methods adopted in this study and an outline of the empirical evaluation methodology are described; classification results obtained by the investigated approaches are presented in the Results section. An analysis of these results materialises in the Discussion section. The paper concludes with a summary of the main contribution and directions for future research.

Method

In this paper, supervised machine learning approaches are employed for the detection of death certificates with notifiable cancer as the cause of death. Three main variables characterise these approaches: 1) the features extracted from the documents (Automatic Feature Extraction), 2) the weighting schemas applied to the features to represent documents (Feature Weighting), and 3) the specific binary classifier used to individuate certificates with notifiable cancer as the cause of death (Automatic Classification). A keyword spotter is described in Baseline and provides a reference for evaluation of the machine learning approaches. The Method section closes with a description of the data, and finally the evaluation strategy.

Automatic Feature Extraction

Machine learning algorithms require data to be represented by features, such as the words that occur in a text document. In this work, the information extraction capabilities of the Medtex system[2] were used for obtaining a set of meaningful features from the free-text of the death certificates shown in Table 1.

Table 1. Feature sets extracted from death certificates by the Medtex system.

Feature	Description
stem	a token stem, i.e. the stemmed version of a word contained in death certificates
conceptFull	the tokens of the fully specified name of the extracted SNOMED CT concepts
concFullBigram	the bigram formed by two adjacent tokens in the fully specified name of concepts extracted from SNOMED CT
stemBigram	the bi-gram formed by two token stems, i.e. a pair of adjacent stemmed words as found in death certificates
concept	SNOMED CT concepts identified in the free-text of the certificates using the Medtex system
concBigram	the bigram formed by two adjacent SNOMED CT concept ids
concFullMorph	the tokens of the fully specified name of extracted SNOMED CT concepts that are morphologic abnormalities or disorders

While features like **stem** and **stemBigram** commonly classify free-text documents, previous research has not considered characteristics based on SNOMED CT concepts and their properties, such as tokens from the fully specified name. SNOMED CT provides a standard clinical terminology used to map various descriptions of a clinical concept to a single standard clinical concept. In this work, the SNOMED CT ontology is used as the underlying mechanism to classify free-text using semantically matching SNOMED CT concepts.

In addition, pair-wise combination of features that showed promising results on preliminary experiments were considered. Results reported for all features used singularly (except for **concFullMorph**), and for the combinations **concept** + **stem**, **concept** + **stemBigram**, **concFullMorph** + **stemBigram**, and **concBigram** + **stemBigram**, showed promise in initial exploration.

Next, consider the example death certificates given in Figure 1 and Figure 2 for describing the composition of a feature set. To build the feature representations, each death certificate is examined, and a value of 1 is assigned for each time a feature occurred in a certificate; while, the absence of a feature receives a value of zero. Note that these values are subsequently modified according to the feature weighting functions, as we shall describe in Feature Weighting. After all certificates have been processed in this manner, a final feature **cancerNotifiable** is added, whose value is obtained from ground truth judgements supplied with the data. Table 2 shows an extract of the feature data

[2] Medtex comprises both information extraction capabilities (extracting both low level information such as word tokens and stems, punctuation, etc., and higher level semantic information such as UMLS and SNOMED CT concepts) and classification capabilities integrated via its rule-based engine (2).

constructed for the two example death certificates. The task of the machine learning classifiers is to predict the value of the **cancerNotifiable** feature, given the learning data supplied.

Note that the text receives no further processing, for example, for removing punctuation, identifying section or list labels, or for removing or correcting typographical errors present in the free-text. While adequate text pre-processing may enhance the quality of the text itself and thus the extracted features, this is left for future research. Instead, weighting schemas for the selected features and binary classifiers are investigated next.

Feature Weighting

A number of weighting schemes for capturing the local importance of a feature in a report were investigated. In the first scheme, Binary coefficients encoded the presence or absence of a feature. This schema is referred to as **binary**.

The weighting schema composed by the feature frequency $f(F)$ of feature F captured the number of times a specific feature appeared within a document. This schema is referred to as **freq**.

Variations of the frequency weighting schema were investigated. These schema directly translate feature frequencies into weights, i.e. weights linearly derived from frequencies. Other variations considered non-linear functions of the frequency of a feature.

A first variation involved scaling the appearance of feature F in a free-text death certificate by the function $1+\log(f(F))$ if f(F)=1, and 0 when absent. This function captures the fact that subsequent appearances of a feature F in a document receive little importance: the logarithm of a number greater than one plateaus rapidly. This schema is referred to as **logF**, i.e. logarithm of the frequency.

A second variation encompasses assigning increasing weights to features that appear with high frequencies within the death certificate. To this aim, the appearance of feature F is weighted according to the function $e^{f(F)}$, while absent features are assigned the value zero. It is suggested that, given the short length of the death certificates under consideration, the unexpected multiple occurrence of a feature would provide strong evidence that feature is important for the document. Using the exponential function to weight occurrences of a feature assigns dominating scores to features that occur frequently in a document. This schema is referred to as **expF**.

Note that scores were assigned to features by using only local weighting functions, that is, weights were computed only by taking into account the frequencies of appearance of a feature within a single text. Thus, the distribution of a feature on a global level, i.e. across the dataset, was ignored when computing the feature weight. The incorporation of global occurrence statistics within the weighting schemas was left for future research.

Automatic Classification

A number of common classifiers were evaluated in this study. These comprised statistical models (Naive Bayes), support vector machines (SPegasos), decision trees (C4.5), and boosting algorithms (AdaBoost). The implementations of these algorithms provided in the Weka toolkit were used in this work (9).

The multinomial Naive Bayes classifier determines the class of a death certificate according to the independent occurrence of features in the text and their weights. The SPegasos classifier uses a stochastic gradient descent algorithm and a hinge loss function to produce the separation hyperplane used by the linear support vector machine. In the C4.5 classifier, information gain provides the choice at each level of the decision tree for the most effective feature able to split the data into the two binary classes considered here (i.e. cancer-related and not cancer-related death certificates). AdaBoost minimises a convex loss function built from the prediction of a base weak classifier. The investigators utilized a simple one-level binary decision tree as the base classifier for AdaBoost. Parameters of all classifiers were set to the default values described in Witten et al (9).

Baseline

A simple keyword spotting classifier was used as a baseline method for the evaluation of the machine learning approaches. The creation of the keyword list came from automatically extracting the set of unique terms from the long description of all ICD-10 cancer codes, and then manually filtering out the non cancer-specific words. A death certificate classification refers to cancer notifiable if it contains one or more terms from the keyword list, otherwise the classification becomes not notifiable.

Data

A set of 5,000 free-text death certificates was acquired from Cancer Institute NSW, the institutional entity responsible for maintaining the Central Cancer Registry in New South Wales. Ethics approval came from the NSW Population & Health Services Research Ethics Committee for this study including use of the de-identified data. The free-text death certificates were short in length, containing on average 13.08 words; the (unstemmed) vocabulary contained 3,751 unique words (including section headings and labels).

Cause of death classifications based on ICD-9 and ICD-10 codes accompanied the reports. This coding set came from the Australian Bureau of Statistics, who release coded data yearly. These ICD coding determined the class to which each death certificate belonged to. The list of notifiable cancer ICD codes was obtained from the neoplasms table in chapter 2 of the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10), version for 2010[3].

The 5,000 death certificates extracted from Cancer Institute

[3] http://apps.who.int/classifications/icd10/browse/en#/II

NSW archives such that 3,111 (62.2%) certificates were coded with ICD codes from the cancer notifiable list according to the neoplasms table. The remaining 1889 (37.8%) contained non-cancer notifiable ICD codes. The coded cause of death for the 3,111 cancer notifiable death certificates represents more than 370 unique ICD cancer related codes.

Evaluation

Many strategies exist that can be employed in the evaluation of machine learning approaches for classification of cancer notifiable death certificates. A 10-fold cross validation becomes applicable in this scenario as the large training set, small test set split at each iteration reflects accurately the application domain, since Cancer Institute NSW possess a large database of coded death certificates from which to build a production quality classification system.

The 10-fold cross validation evaluation strategy was thus used to train and test the classifiers. In this methodology, the dataset is randomly divided into 10 stratified[4] folds of equal dimensions. A model for each classifier is then learnt on nine of these folds, leaving one fold out for testing. The process is repeated by selecting a new fold for testing, while a new model is learnt from the remaining folds. The classifier constructed at each iteration is run against the fold left out for testing and the result is recorded. Once all folds are complete, classification effectiveness is averaged across all test fold results, and this measure is reported herein.

F-Measure (F-m) was used as the primary metric to evaluate the efficacy of the implemented classifiers; accuracy, recall (sensitivity, Rec) and precision (positive predictive value, Prec) were also recorded, along with the number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) classifications.

Results

The combination of 10 features, four weighting schemas, and four classifiers requires the evaluation of a total of 160 classifier settings (referred to as runs in the following) on the dataset consisting of 5,000 death certificates. While all combinations of features, weighting schema and classifiers were evaluated, due to the large number of combinations it is unfeasible to report the individual results for each of the runs. Thus, only the settings of the 40 most effective runs in terms of F-measure, the primary evaluation metric (Table 3), are reported.

The F-measure of each classifier over all experimented settings is graphically shown in Figure 3. Later in the paper, a summary evaluation of the variability of results provided by features, weighting schemas, and classifiers is reported. This analysis considers the results from all runs.

[4] Folds were automatically stratified with respect to the two target classes, not the ICD-10 codes.

The results reported in Table 3 offer strong suggestion that the tested approaches produce highly effective results in discriminating between those death certificates that contain a cancer notifiable cause of death and those death certificates that contain a non-cancer notifiable cause of death. A reference comparison can be made with the performance of the baseline keyword classifier presented in Table 4. Overall, the support vector machine implementation provided by SPegasos emerged as the best classifier when used on the **stem** feature set, weighted using binary coefficient. SPegasos happens to be very effective also when other combinations of weighting schemas and features come under consideration, accounting for 20 of the presented top 40 results, of which it occupies the first 18 positions. Additionally, the SPegasos classifier shows lower variance across all settings compared to Naive Bayes, C4.5 and Adaboost, being 47.7%, 61% and 61.9% lower respectively (Figure 3 & Table 5).

The keyword spotting baseline classifier produced an F-measure of 0.8837, precision of 0.9736 and recall of 0.8071 (Table 4). The top 40 machine learning classifiers presented here have 7-10% higher F-measure than the baseline classifier.

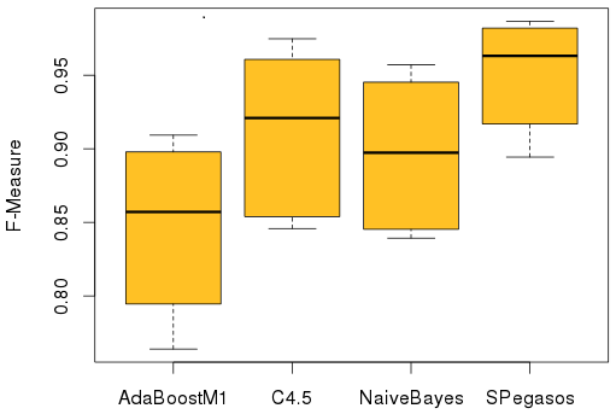
Discussion

To better understand the role of specific feature sets, weighting schema and classifiers on the effectiveness of the tested approaches, we performed an analysis of the empirical results with independent designation of each of the three key characteristics as controlled variables.

Table 4. Baseline keyword classifier result.

Prec	Rec	F-m	TP	FN	FP	TN
0.9763	0.8071	0.8837	2511	600	61	1828

Figure 3. Boxplot summarising F-measure of the investigated classifiers over all considered settings.



Classifiers

We start by examining the impact of each classification model on the overall effectiveness of the approaches. Table 5 reports maximum (Max), minimum (Min), difference (Δ), and variance (Var) of F-measure over all runs of each classifier model. SPegasos appeared as the classifier achieving the highest maximum F-measure (0.9866), highest minimum F-measure (0.8944), lowest difference (0.0922)

and lowest variance (1.14E-03), extending the observations made on this classifier when examining the results of Table 3; all SPegasos results emerged with higher readings than the keyword classifier baseline. The Naive Bayes classifier failed to be amongst the most effective classification models in our experiments; however, its robustness comes second only to that of SPegasos, with difference and variance of only 0.1178 and 2.18E-03 in F-Measure respectively. A classifier that exhibits low values of difference and variance across its different settings (i.e. weighting schema, feature sets, etc.) indicates it may be less susceptible to variances in effectiveness when applied to unseen data. In our experiments, this is the case for SPegasos and Naive Bayes. However, while C4.5 achieves higher values of F-measure than Naive Bayes, its difference and variance are greater. AdaBoost showed the worst performance overall with 28 of its 40 runs performing worse than the baseline.

Table 5. Classification effectiveness of all four classifiers using F-measure, ordered by increasing variance.

Classifier	Max	Min	Δ	Var
SPegasos	0.9866	0.8944	0.0922	1.14E-03
Naive Bayes	0.9571	0.8393	0.1178	2.18E-03
C4.5	0.9748	0.8458	0.129	2.92E-03
AdaBoost M1	0.9093	0.7639	0.1454	2.99E-03

Weighting Schema

We continue by analysing the influence of weighting schemas on the classification effectiveness (see Table 6). The simple binary coefficient schema achieved the highest F-measure. However, no weighting schema appears to be significantly better than another: while **binary** achieves the best performance with an F-measure of 0.9866, the highest F-measure of the worst performing schema, **expF**, came to 0.9821; just 0.46% lower than **binary**. Furthermore, all weighting schema exhibit the same effectiveness when considering the worst performing settings. Thus the range of performance differences and their variance do not significantly differ across weighting schema. Weighting schema shows the highest average variance at 3.77E-03, in comparison to Classifier and Feature set average variance of 2.31E-03 and 1.70E-03 respectively. This may be due to the fact that death certificates appear in general short documents, where features occur uniformly and thus different schemas to produce feature weights do not sensibly differ.

Table 6. Classification effectiveness across the four weighting schema using F-measure, ordered by increasing variance.

Weight	Max	Min	Δ	Var
expF	0.9821	0.7639	0.2182	3.58E-03
logF	0.9858	0.7639	0.2219	3.77E-03
freq	0.9848	0.7639	0.2209	3.82E-03
binary	0.9866	0.7639	0.2227	3.89E-03

Feature Sets

Feature set is the final variable controlled for in our analysis, and arguably the one with the greatest impact on classification results (Table 7). The use of the **stem** feature set provides the highest F-measure (0.9866), while **concBigram** yields the lowest maximal F-measure (0.9171): a marked difference of 7.03%. The **concept** + **stem** (9.82E-04) demonstrated the smallest variance, making it the most robust feature set in our experiment; in addition this feature yielded a maximal F-measure only 0.02% lower than the best value recorded in our experiments. The average variance of feature sets appeared as the lowest of the controlled variables at 1.70E-03, as noted previously. These results provide strong indication that, of the variables analysed, the choice of feature provides the greatest contribution to the classification effectiveness, although careful selection of the classifier can provide significant improvement. To illustrate the importance of feature set selection, when the **concept** + **stem** feature set undergoes combination with the worst performing classifier in our experiment, **AdaBoost**, F-measure of the pair becomes 0.9093, the highest recorded for **AdaBoost**.

Related Work

The results presented here showed improved performances over an earlier publication (10). Previous work considered the NSW Cancer Registry (NSW CR) business rules, which narrow the conditions for notifiable death certificates to both a) cancer related cause of death and b) issued for a NSW CR patient. The current work considers the more general rules applied by the Australian Bureau of Statistics (ABS), which only require that the death certificate possesses a cancer related cause of death. This study also extends the scope of the work to use ABS ICD-9 and ICD-10 coded death certificates.

Table 7. Classification effectiveness across all ten feature sets using F-measure, ordered by increasing variance.

Feature	Max	Min	Δ	Var
concept + stem	0.9864	0.9085	0.0779	9.82E-04
concept + stemBigram	0.9825	0.8979	0.0846	1.04E-03
stem	0.9866	0.901	0.0856	1.05E-03
concFullBigram	0.895	0.7945	0.1005	1.35E-03
conceptFull	0.9254	0.8266	0.0988	1.36E-03
stemBigram	0.9801	0.8767	0.1034	1.59E-03
concFullMorph + stemBigram	0.9825	0.8766	0.1059	1.62E-03
concept	0.9526	0.8376	0.115	1.77E-03
concBigram + stemBigram	0.9171	0.7639	0.1532	3.13E-03
concBigram	0.9171	0.7639	0.1532	3.13E-03

In this study, the use of the ABS rules is found to improve performance over that of the NSW CR rules used in the previous study (10). The difference in specificity between the ABS and NSW CR rules; the extended scope of this study; and the analysis of classifier predictions being applied

to this study were new contributions from this current study.

Conclusion

Timely processing of cancer notifications remains critical for timely reporting of cancer incidence and mortality. Death certificates create a rich source of data on cancer mortality. Cancer registries acquire free-text death certificates on a regular (e.g. fortnightly) basis. However, the cause of death information needs to be classified to facilitate reporting of cancer mortality. Cause of death information classified using ICD codes becomes available only on an annual basis. In this paper, the automatic classification of death certificates was studied to individuate cancer notifiable cause of death. The investigated approaches achieved overall strong classification effectiveness, with a support vector machine classifier trained with token stem features and weighted by a simple binary coefficient of appearance in the document yielding an F-measure of 0.9866. The choice of feature set, and of classifier, represented determining factors for high effectiveness. The weighting schema had no appreciable effect on classification effectiveness. The use of an automatic classification system of similar effectiveness to the description presented here possesses the potential to improve workflows for the coding of cancer notifiable free-text death certificates.

Future efforts need to be directed towards an in depth error analysis, in particular examining the distance between the prediction produced by a classifier and the decision threshold. Further work will be directed towards extending the current methods to predict the actual ICD-10 codes associated with a cause of death related to cancer, so as to further assist clinical coders in processing cancer notifications.

References

1. Coleman MP, Forman D, Bryant H, Butler J, Rachet B, Margine C, Nur U, Tracey E, Coory M, Hatcher J, McGahan CE, Turner D, Marret L, Gjerstorff ML, Johannesen TB, Adolfsson J, Lambe M, Lawrence G, Meechan D, Morris EJ, Middleton R, Steward J, Richards MA, ICBP Module 1 Working Group. 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. The Lancet. 2011;377(9760);127-138.

2. Nguyen A, Moore J, Lawley M, Hansen D, Colquist S. Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications. Health Inform Conference. 2011;117–124.

3. Zuccon G, Nguyen A, Bergheim A, Wickman S, Grayson N. The impact of OCR accuracy on automated cancer classification of pathology reports. Stud Health Technol Inform. 2012;178;250.

4. D'Avolio L, Nguyen T, Farwell W, Chen Y, Fitzmeyer F, Harris O, Fiore L. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). J Am Med Inform Assoc. 2010;17(4);375–382.

5. Harris, K. Selected data editing procedures in an automated multiple cause of death coding system. Proc Conference Eur Stat. 1999.

6. Davis K, Staes C, Duncan J, Igo S, Facelli J. Identification of pneumonia and influenza deaths using the death certificate pipeline. BMC Med Infor Decis Making. 2012;12(1);37.

7. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, Colquist S. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc. 2010;17(4);440–445.

8. Stearns MQ et al. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp J Am Med Inform Assoc. 2001.

9. Witten I, Frank E, Hall M. Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques. Morgan Kaufmann; 2011.

10. Butt L, Zuccon G, Nguyen AN, Bergheim A, Grayson N. Automatic Classification of Cancer Notifiable Death Certificates. CEUR Workshop Proc. 2012;941;65-76.

ACKNOWLEDGEMENTS

The authors would like to thank the Australian e-Health Research Centre and Cancer Institute NSW for their support of this research. The authors also recognise the work of the ABS and NSW CR in collecting and coding death certificates, without which this work would have been possible.

PEER REVIEW

Not commissioned. Externally peer reviewed.

CONFLICTS OF INTEREST

The authors declare that they have no competing interests.

FUNDING

Internally funded by AEHRC and CINSW.

ETHICS COMMITTEE APPROVAL

NSW Population & Health Services Research Ethics Committee, HREC/11/CIPHS/60.

Table 2. Feature data built from two example death certificates.

	Features																			
	stem						stemBigram					concept			conceptFull					
Document	ACCID	ALCOHOL	...	TUMOR	WEEK	YEAR	ACCID_DYSPLASIA	ACCID_48	...	20_YEAR	YEAR_ASTHMA	12655004	...	230690007	Neoplasm of maxilla	...	Cerebrovascular accident	Cerebral arteriosclerosis	...	cancerNotifiable
Figure 1	1	0	...	1	1	1	1	0	...	1	1	1	...	1	1	...	1	0	...	1
Figure 2	1	1	...	0	0	1	0	1	...	0	0	0	...	1	0	...	1	1	...	0

Table 3. Top 40 results with respect to decreasing F-measure (F-m).

Classifier	Feature	Weight	Prec	Recall	F-m	TP	FN	FP	TN
SPegasos	stem	binary	0.9922	0.981	0.9866	3052	59	24	1865
SPegasos	concept + stem	binary	0.9912	0.9817	0.9864	3054	57	27	1862
SPegasos	stem	logF	0.99	0.9817	0.9858	3054	57	31	1858
SPegasos	concept + stem	logF	0.9922	0.9791	0.9856	3046	65	24	1865
SPegasos	stem	freq	0.989	0.9807	0.9848	3051	60	34	1855
SPegasos	concept + stem	freq	0.9889	0.9775	0.9832	3041	70	34	1855
SPegasos	concept + stemBigram	logF	0.9915	0.9736	0.9825	3029	82	26	1863
SPegasos	concFullMorph + stemBigram	freq	0.9892	0.9759	0.9825	3036	75	33	1856
SPegasos	concept + stemBigram	binary	0.9925	0.9724	0.9823	3025	86	23	1866
SPegasos	concFullMorph + stemBigram	expF	0.9876	0.9765	0.9821	3038	73	38	1851
SPegasos	concFullMorph + stemBigram	logF	0.9889	0.9752	0.982	3034	77	34	1855
SPegasos	concept + stemBigram	freq	0.9896	0.9743	0.9819	3031	80	32	1857
SPegasos	concept + stemBigram	expF	0.9886	0.9749	0.9817	3033	78	35	1854
SPegasos	concFullMorph + stemBigram	binary	0.9895	0.9724	0.9809	3025	86	32	1857
SPegasos	stemBigram	freq	0.9882	0.972	0.9801	3024	87	36	1853
SPegasos	stemBigram	binary	0.9895	0.9701	0.9797	3018	93	32	1857
SPegasos	stemBigram	logF	0.9892	0.9704	0.9797	3019	92	33	1856
SPegasos	stemBigram	expF	0.9879	0.9717	0.9797	3023	88	37	1852
C4.5	concept + stem	binary	0.9814	0.9682	0.9748	3012	99	57	1832
SPegasos	stem	expF	0.9827	0.9666	0.9746	3007	104	53	1836
C4.5	stem	binary	0.9849	0.964	0.9743	2999	112	46	1843
SPegasos	concept + stem	expF	0.9792	0.9682	0.9737	3012	99	64	1825
C4.5	stem	logF	0.9808	0.9666	0.9736	3007	104	59	1830
C4.5	stem	expF	0.9808	0.9666	0.9736	3007	104	59	1830
C4.5	stem	freq	0.9808	0.9666	0.9736	3007	104	59	1830
C4.5	concept + stem	logF	0.9798	0.965	0.9723	3002	109	62	1827
C4.5	concept + stem	expF	0.9798	0.965	0.9723	3002	109	62	1827
C4.5	concept + stem	freq	0.9798	0.965	0.9723	3002	109	62	1827
C4.5	concept + stemBigram	logF	0.9647	0.9569	0.9608	2977	134	109	1780
C4.5	concept + stemBigram	expF	0.9647	0.9569	0.9608	2977	134	109	1780
C4.5	concept + stemBigram	freq	0.9647	0.9569	0.9608	2977	134	109	1780
C4.5	concept + stemBigram	binary	0.9662	0.955	0.9606	2971	140	104	1785
Naive Bayes	concept + stem	logF	0.9569	0.9572	0.9571	2978	133	134	1755
Naive Bayes	stem	logF	0.9599	0.9544	0.9571	2969	142	124	1765
C4.5	stemBigram	binary	0.9577	0.9531	0.9554	2965	146	131	1758
C4.5	stemBigram	logF	0.9577	0.9527	0.9552	2964	147	131	1758
C4.5	stemBigram	expF	0.9577	0.9527	0.9552	2964	147	131	1758
C4.5	stemBigram	freq	0.9577	0.9527	0.9552	2964	147	131	1758
C4.5	concFullMorph + stemBigram	binary	0.9556	0.9544	0.955	2969	142	138	1751