Extracting Cancer Mortality Statistics from Free-text Death Certificates

A View from the Trenches

Bevan Koopman[1], Anthony Nguyen[1], Danica Cossio[2], Mary-Jane Courage[2], Gary Francois[2]
[1]Australian e-Health Research Centre, CSIRO
[2]Queensland Cancer Control Analysis Team, Queensland Health
Brisbane, Australia

ABSTRACT

This industry paper describes a deep learning and information retrieval system that allows cancer registries to extract cancer mortality statistics from free-text death certificates. Death certificates may provide an invaluable source of mortality information but to realise this value automated methods for classifying cancer types and searching certificates are needed. We present a system comprising a deep learning classifier to identify cancer related deaths, an IR system to allow users to search death certificates and classifier results, and a deployment architecture that aims to handle issues of scalability and complexity. Empirically, the system can accurately identify cancer deaths for both common and rare cancers. The use of the IR system helps users drill into specific results and convince them of the utility of using an automated approach. The paper aims to touch on a number of issues in applying deep learning and IR techniques to real-world settings.

KEYWORDS

Cancer mortality, Death certificates

ACM Reference Format:
Bevan Koopman[1], Anthony Nguyen[1], Danica Cossio[2], Mary-Jane Courage[2], Gary Francois[2]. 2018. Extracting Cancer Mortality Statistics from Free-text Death Certificates: A View from the Trenches. In *23rd Australasian Document Computing Symposium (ADCS '18), December 11–12, 2018, Dunedin, New Zealand.* ACM, New York, NY, USA, Article 4, 5 pages. https://doi.org/10.1145/3291992.3292003

1 INTRODUCTION

A death certificate is a legal document, issued by a medical practitioner, certifying the cause of a person's death. In aggregate, the cause-of-death represents a vital source of mortality statistics [1]. Cancer registries[1] rely heavily on such data to provide an accurate picture of the impact of cancer, the effect of cancer treatments and to direct research efforts for cancer control. However, cancer registries receive an overwhelming number of free-text death certificates; each of which needs to be manually assessed to determine if it is a cancer related death and, if so, the specific type of cancer. The problem is compounded by the fact that some cancers are extremely rare with only a few cases a year, while others are very common.

```
(A) LIVER FAILURE (B) LIVER METASTASES (C) BREAST CANCER
```

Figure 1: Sample death certificate. The underlying cause-of-death is breast cancer (C50).

A sample, cancer-related, death certificate is shown in Figure 1. Death certificates are authored according to a specific procedure: Statement A) being the "Disease or condition directly leading to death" and the ordering interpreted as A) due to B) due to C), and C) being the *underlying* cause of death.

In this paper, we describe a system to automatically extract cancer statistics from death certificates. The system is comprised of: 1) a deep learning system that, given a death certificate, assigns an appropriate cancer class according to ICD10[2]; 2) an IR system with web UI allowing users to search death certificates using both ad-hoc queries and ICD10 codes; and 3) an integrated system architecture with docker-based micro-services for flexible and scalable deployment of the system within cancer registries. This paper describes an industry project between researchers at CSIRO and Queensland Cancer Control Analysis Team within Queensland Health.

Existing approaches to classifying death certificates have been either rule-based [4, 5] or machine learning (ML) [2]. While ML was generally more effective they had performed very poorly for rare cancers. They also used extensive domain specific features from an NLP pipeline and implemented multiple binary classifiers for each cancer - the scalability and complexity of such a system was prohibitive. To overcome the poor performance of rare cancers a hybrid rule/ML approach is possible [3]. While this solved the problem of rare cancers it added an extra layer of complexity of integrating rules and ML models and how to normalise classification scores between the two which risked overfitting. It also resulted in a complex, resource intensive deployment architecture. While these existing approaches do perform quite well empirically, their ability to translate from the lab to a production environment is hampered by issues of complexity and scalability, both in training the models and in running when in production.

The contribution of this paper is a system that overcame a number of these issues. The models used simple term features, alleviating the need for domain-specific NLP pipelines. A single multi-class, rather than multiple binary, model was developed (alleviating score normalisation between models). A lightweight resource footprint

[1]Cancer registries are organisations responsible for the reporting and monitoring of cancer in the general population.

[2]ICD10 is a medical classification system that contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases.

ADCS '18, December 11–12, 2018, Dunedin, New Zealand

ACM ISBN 978-1-4503-6549-9/18/12.
https://doi.org/10.1145/3291992.3292003

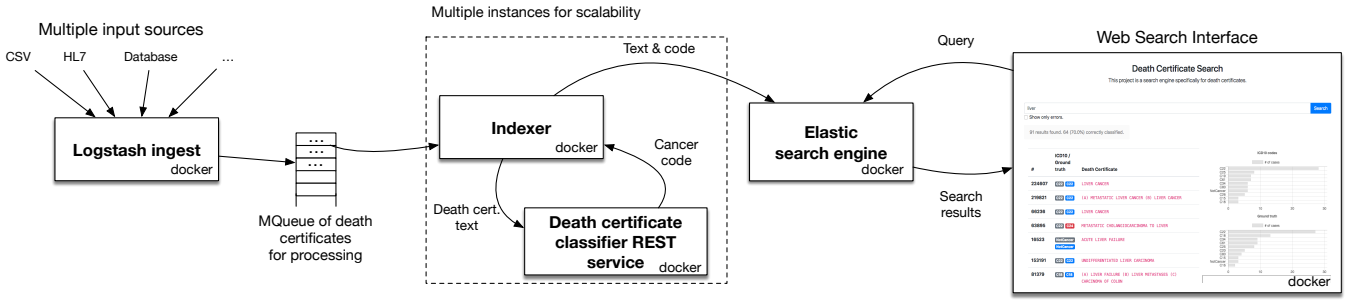

Figure 2: System architecture (read left to right). Death certificates were input from multiple sources. They were classified and then indexed into an Elastic search engine. Ad-hoc search and result summarisation was provided via a web interface.

(CPU & memory), as well as flexible deployment architecture, made a more scalable and efficient system. Our empirical evaluation on one year's worth of certificates showed that the system could accurately classify most cancers (both common and rare). Analysis of errors showed that these were often where the death certificate did not accurately reflect the cause of death (e.g., where later investigation contradicted the initial cause of death).

2 METHODS

2.1 Death certificate dataset & ground truth

A collection of 355,164 death certificates, covering all deaths between 1999 and 2006 was used for training. A key characteristic of this dataset was that each contained a cause-of-death code provided by the Australian Bureau of Statistics. Therefore, cancer codes represented the ground truth for training a deep learning model for automated classification of cancer deaths.

The use case for the proposed system was the Queensland Cancer Control Analysis Team within Queensland Health. They provided 29,560 certificates covering all deaths for 2015. At the time, these were manually coded with cancer related cause-of-death-codes. This represented the validation set to evaluate the proposed system.

2.2 Cancer classification

Given a free-text death certificate, the aim was to 1) identify if the death was cancer related and; 2) determine the specific type of cancer (according to the ICD10 medical classification system). To this end, we trained a multilayer perceptron deep learning-based model. This was a multi-class classifier with each class representing a cancer related ICD10 code and additional class of "NotCancer" for all other non-cancer deaths. As features, we used word tokens, represented as one-hot encoded binary vectors. The architecture had 3 layers: input layer with a size amounting to the vocabulary size (i.e., number of unique words in the collection), a hidden layer of size 500 nodes, and output layer of size 71, representing each of the different applicable cancer types.

Training was done using a 75% / 25% train/validation split of the test collection with 5 epochs. On completion, the model was serialised for deployment within a wider deployment architecture we detail later.

Two requirements existed for cancer identification: identify if the cause of death was cancer and determine the type of cancer. The former is a binary cancer/no-cancer classification; the later a multi-class ICD10 classification. The binary classification was determined via a possible classification for any of the ICD10 cancer codes; negative classification was taken as the "NotCancer" label.

The classifier was implemented in Python's Keras package using the Tensorflow library.

A hybrid rule/ML approach is chosen as a benchmark for comparison [3]. The same training and test collections as [3] were used to evaluate the deep learning methods outlined in the paper.

2.3 Search engine

While the cancer classification provided by the deep learning model was important, the overall value of such a system was only realised if humans can easily interact with the data. Toward this aim, a search engine was implemented allowing users to search both the free-text of death certificates, the cancer classification of the model and the ground truth (where available). The purpose of the search engine was threefold: 1) Provide a means for users — typically analysts from the Cancer Control Analysis Team — to issue ad-hoc queries across the collection of death certificates to investigate specific cancers or conditions. Users would like to both see a ranked list of relevant death certificates to their query, as well as summary statistics of the number of each cancer types found in the set of results. 2) Provide a means for users to understand and monitor the performance of the automated classification system, assuring them of its effectiveness before its fully adopted. 3) Identify individual cases where cancer related deaths may have been misdiagnosed.

All death certificates were indexed in a fielded Elasticsearch instance; fields included the original death certificate text, the ICD10 classification from the classifier, the description of the ICD10 code (e.g., "Lung cancer" for C43), and the ground truth code (if available). This allowed user to search both text and ICD10 codes (which users were very familiar with). Ranking was done in BM25.

A web-based interface was implemented with a single text box for ad-hoc queries. Given a query, a ranked list of death certificates was provided, including the cancer classification and ground truth (if available). Summary histograms showing a breakdown of the cancer types were also shown. (More details on the search interface are provided in the results section.)

2.4 Deployment Architecture

Figure 2 provides an overview of the system architecture. To support a decoupled and scalable architecture, individual components were decoupled and deployed as separate docker containers. From

Precision	29,560		**Classifier**	
0.8740			-	+
Recall	**Ground**	-	19,899	1,176
0.9720	**truth**	+	236	8,159

Table 1: Binary classification results with confusion matrix; + denotes *cancer* and - denotes *nocancer*.

	Precision	**Recall**
Micro avg.	0.8726	0.9675
Macro avg.	0.8408	0.9303

Table 2: Average classification results across all classes.

	Precision	**Recall**
Hydrid rule/ML benchmark [3]	0.8086	0.8054
Proposed	0.8408	0.9303

Table 3: Comparison against hybrid rule/ML benchmark [3].

left to right in the figure: 1) A **Logstash**[3] instance ingested death certificates from any source type that has Logstash support. This could be a single certificate from a daily trickle feed, for example, or a batch archived from previous years and stored in a database. Logstash placed each certificate on a queue ready for processing. 2) The **Indexer** popped items off the queue and issued the death certificate to a standalone **Death certificate classifier** via an HTTP REST interface. The resulting classification (and probability representing confidence) was returned. To support scalable processing of many certificates, multiple Indexers & Death certificate classifiers could be run in parallel. 3) The death certificate and classification (ICD10 code and cancer description) were indexed into the **Elastic search engine**. 4) Users interacted via a Web search interface that talked to the **Elastic search engine**.

3 RESULTS & ANALYSIS

3.1 Classification results

The binary (cancer or no cancer) classification results are shown in Table 1, including precision, recall and a confusion matrix. Recall was high, indicating the classifier found most cancer related deaths. Precision was still high but reduced somewhat by false positives. Indeed, from the confusion matrix, false positives made up the majority of errors. For this task, a false negative (i.e., missed cancer death) was more harmful than a false positive; thus the systems higher recall at the expense of precision was desirable.

Individual cancer results are shown in Figure 3. Results are ordered by prevalence (common cancers on top; rare cancer at the bottom). The data shows that in Queensland the five most common cancers deaths were lung, colon, prostate, breast and pancreatic. Recall was high for all cancer but, more importantly, remained high for rare cancers. A limitation here being that for rare cancers there were only a few samples for evaluation. Lower precision (i.e., more false positives) was observed mainly for cancers which were not very common or very rare (i.e., in the middle).

To provide an overall estimate of cancer type classification effectiveness, Table 2 reports micro and macro average precision and recall. In addition, comparison with the hybrid rule/ML benchmark of [3] are shown in Table 3. The less complex method outlined in this paper outperformed the existing benchmark.

[3]Logstash is a data processing pipeline that allows for the collection of data from a variety of sources, transforming it on the fly, and sending it to some desired destination.

Figure 3: Classification results for individual cancers.

3.2 Search engine

A screenshot of the web search interface is shown in Figure 4. A table of results on the left of the screen showed death certificates matching the query "kidney". For each result, the death certificate text was shown as well as a grey label indicating the ICD10 classification for that certificate. If ground truth data was available then the ICD10 ground truth code was shown: blue labels indicated the classification matched the ground truth; red label indicates an incorrect classification. A mouse hover over any ICD10 label provided a black popup with the description of the cancer code and the probability of that classification.

Two histograms were displayed on the right of the screen; they showed the distributions of cancers for the current set of search

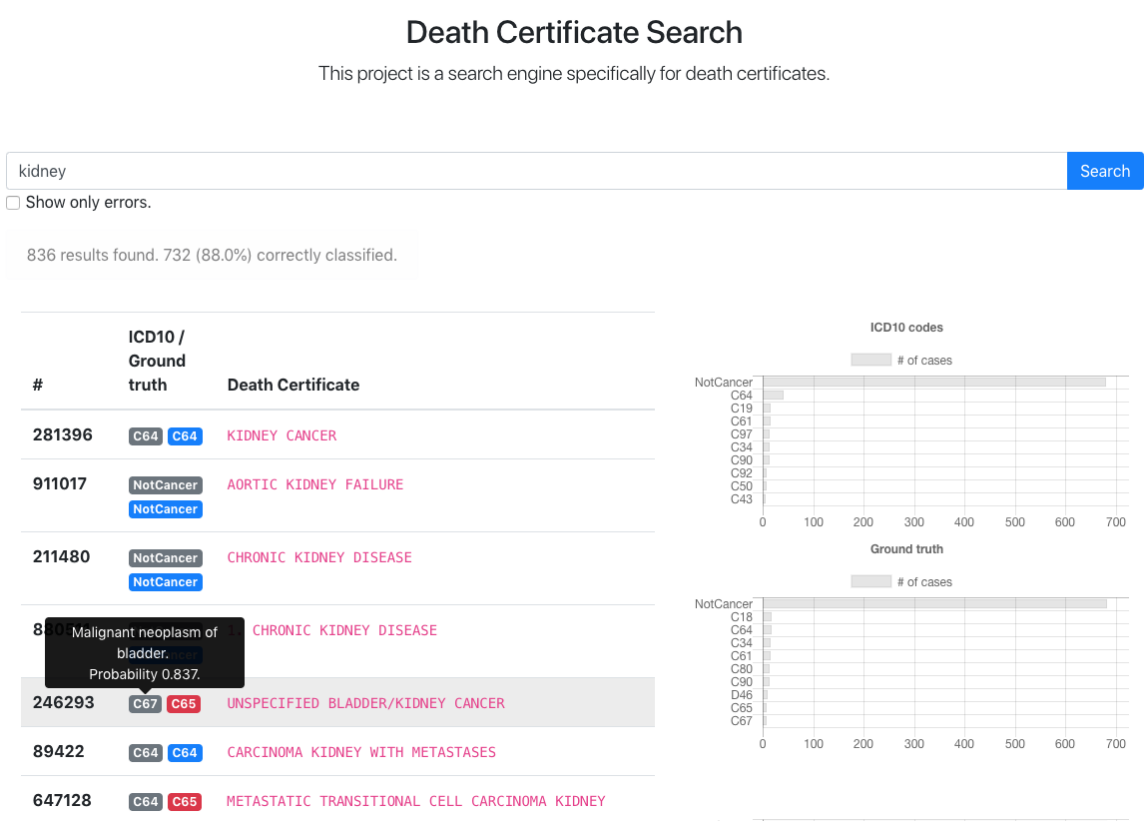

Figure 4: Web search interface.

results (for both the classifier and the ground truth data). This provided the user with a simple visual overview of results. It also allowed for a quick understanding of incorrect classifications. Before putting the system into production, users needed to both be confident that the system was effective and to understand the impact of errors. This interface provided a good mechanism to drill into results and allow the users to assess the utility of the system.

Initial use of the search system by end-users has revealed: 1) Difficulties in determining the underlying cause of death where there were multiple conditions. For example, the certificate: Breast NotCancer `(A) MYOCARDIAL INFARCT (B) CARCINOMA BREAST`. The classifier indicated breast cancer but the cause of death was not cancer; instead it was a heart attack (also known as myocardial infarct). Further investigation indicated the breast cancer was 10 years ago and hence not related to the heart attack. A number of certificates displayed such characteristics. 2) Situations where the actual cause of death was not known at the time the death certificate was authored. For example, the certificate Brain Lung `METASTASES TO BRAIN`, where the cancer had spread to the brain but actually originated from the lung (only known after the certificate was authored). 3) That skin cancers in the form of squamous cell and basal cell carcinoma were special cases that should not be notified as a cancer case.

4 CONCLUSION

Accurate cancer mortality statistics can be extracted from free-text death certificates with appropriate automated methods. This paper describes a system to do this via three components: a deep learning classifier to determine a specific cancer cause-of-death from a death certificate; an IR system (with web UI) to search the results, study specific cancers and convey to users that the classifier is effective; a scalable deployment architecture that overcomes some of the barriers of putting such systems into production. The classifier achieved a precision and recall of 0.87 and 0.97 respectively — effective enough for adoption. The IR system helped users better understand their death certificates and how they were classified. Future work will look at integrating the system with other sources of data, such as pathology reports, particularly for cases where the cause-of-death is ambiguous or not known from the death certificate. Relevance assessments will also be collected to evaluate and improve the search engine. Lessons from this paper are intended to help those applying deep learning and IR to real-world applications.

REFERENCES

[1] Robert R German, Aliza K Fink, Melonie Heron, Sherri L Stewart, Chris J Johnson, Jack L Finch, and Daixin Yin. 2011. The accuracy of cancer mortality statistics based on death certificates in the United States. *Cancer epidemiology* 35, 2 (2011), 126–131.

[2] Bevan Koopman, Guido Zuccon, Anthony Nguyen, Anton Bergheim, and Narelle Grayson. 2015. Automatic ICD-10 Classification of Cancers from Free-text Death Certificates. *Journal of Medical Informatics* 84, 11 (2015), 956–965.

[3] Bevan Koopman, Guido Zuccon, Anthony Nguyen, Anton Bergheim, and Narelle Grayson. 2018. Extracting Cancer Mortality Statistics from Death Certificates: A Hybrid Machine Learning and Rule-based Approach for Common and Rare Cancers. *Artificial Intelligence in Medicine* To appear (2018).

[4] Bill Riedl, Nhan Than, and Michael Hogarth. 2010. Using the UMLS and Simple Statistical Methods to Semantically Categorize Causes of Death on Death Certificates. In *AMIA Annual Symposium Proceedings*, Vol. 2010. 677.

[5] Anoop D Shah, Carlos Martinez, and Harry Hemingway. 2012. The freetext matching algorithm: a computer program to extract diagnoses and causes of death from unstructured text in electronic health records. *BMC Medical Informatics and Decision Making* 12, 1 (2012), 88.