What Can We Deep Learn About Cancer Mortality?

Bevan KOOPMAN [a], Anthony NGUYEN [a], Danica COSSIO [b], Mary-Jane COURAGE [b], Gary FRANCOIS [b]
[a] *Australian e-Health Research Centre, CSIRO*
[b] *Queensland Cancer Control Analysis Team, Queensland Health*

Introduction

The Queensland Cancer Control & Analysis Team (QCCAT) rely heavily on death certificates (e.g., Figure 1) to provide an accurate picture of the impact of cancer, the effect of cancer treatments and to direct research efforts for cancer control. However, they receive an overwhelming number of free-text death certificates; each needs to be manually assessed to determine if the death is cancer related and the specific type of cancer. To overcome this, automated methods for classifying cancer types and searching certificates are needed.

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

Figure 1: Sample death certificate; breast cancer indicated as underlying cause of death.

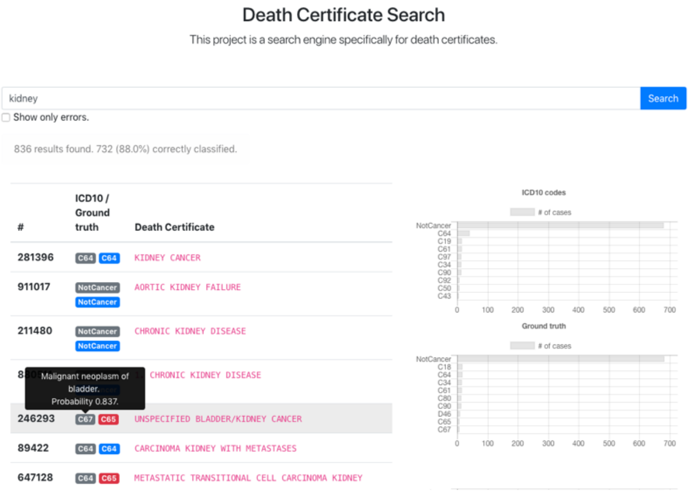
Approach

We developed such an automated system comprising: i) a *deep learning class*ifier to identify cancer related deaths; ii) a *search engine* allowing users to search death certificates and classifier results; and iii) a *deployment architecture* that handles issues of scalability and complexity.

355,164 death certificates, covering all deaths between 1999-2006 [2], were used to train a Tensorflow deep learning classifier. Given a free-text death certificate, the classifier assigned a cancer related ICD10 code. To evaluate the system, a separate 29,560 certificate test set, covering all QLD deaths for 2015, was used.

Body

Classifier accuracy was evaluated to be 92% and, importantly, was effective for both rare and common cancers.



Death Certificate Search

This project is a search engine specifically for death certificates.

kidney [Search]

□ Show only errors.

836 results found. 732 (88.0%) correctly classified.

#	ICD10 / Ground truth	Death Certificate
281396	C64 C64	KIDNEY CANCER
911017	NotCancer NotCancer	AORTIC KIDNEY FAILURE
211480	NotCancer NotCancer	CHRONIC KIDNEY DISEASE
8[illegible]		CHRONIC KIDNEY DISEASE
246293	C67 C65	UNSPECIFIED BLADDER/KIDNEY CANCER
89422	C64 C64	CARCINOMA KIDNEY WITH METASTASES
647128	C64 C65	METASTATIC TRANSITIONAL CELL CARCINOMA KIDNEY

Malignant neoplasm of bladder.
Probability 0.837.

Figure 2: Search Engine for Death Certificates

While classifier accuracy was high, the overall value of such a system was only realised if humans can easily interact with the data. Toward this aim, a search engine (Figure 2) allowed users (analysts from the Cancer Control Analysis Team) to search both the free-text of death certificates and ICD10 codes. The purpose was to provide a means for users to:
1) Investigate specific cancers or conditions by issuing queries across the collection of death certificates. Results comprised individual death certificates and summary statistics in graph form (Figure 2).
2) Understand and monitor the performance of the classifier, satisfying users with its effectiveness prior to adoption.
3) Identify individual cases where cancer related deaths may have been misdiagnosed.

Deployment was done via a scalable and flexible architecture (Figure 3). Individual components were decoupled and deployed as separate docker containers. *Logstash ingest* provided ingestion from various sources (trickle or batch). A processing queue allowed for flexibility and fail-safe processing. *Indexer* issued the death certificate to a standalone *Death certificate classifier REST service*. The resulting ICD10 code was returned and indexed into the Elastic search engine. (Multiple Indexers & Death certificate classifiers could be run in parallel for scalability). Users interacted via a Web search interface (Figure 2).

Through the application of machine learning and search engine technology, an automated system now provides real-time classification of cancer related death and a search system for detailed analysis of results.

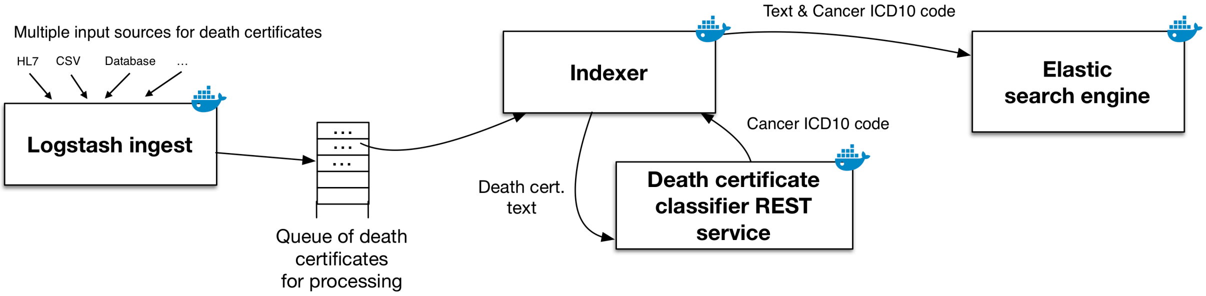

Multiple input sources for death certificates
HL7
CSV
Database
...
Logstash ingest
Queue of death certificates for processing
Indexer
Death cert. text
Death certificate classifier REST service
Cancer ICD10 code
Text & Cancer ICD10 code
Elastic search engine

Figure 3: System architecture.

Conclusions

Accurate cancer mortality statistics can be extracted from free-text death certificates via a system with three components: a deep learning classifier to determine the specific cancer with 92% accuracy; a search 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 search system helped users better understand their death certificates and how they were classified and assure them of its effectiveness before being fully adopted. Future work will look at other sources of data for cancer statistics such as pathology reports.

Keywords

Cancer Mortality Statistics, Death Certificates, Deep Learning

1. Robert German, et al. 2011. **The accuracy of cancer mortality statistics based on death certificates in the United States**. Cancer epidemiology 35, 2 (2011), 126–131.
2. B. Koopman, et al. **Extracting cancer mortality statistics from death certificates: A hybrid machine learning and rule-based approach for common and rare cancers**. Artificial Intelligence in Medicine, 89:1–9, July 2018.