# Automatic ICD-10 Classification of Cancers from Free-text Death Certificates

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*

**Abstract**

**Objective:** Death certificates provide an invaluable source for 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 nature of certificates written in natural language. This paper proposes an automatic classification system for identifying 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. These features were used to train Support Vector Machine classifiers (one classifier for each cancer type). The classifiers were deployed in a cascaded architecture: the first level identified the presence of cancer (i.e., binary cancer/no cancer) and the second level identified the type of cancer (according to the ICD-10 classification system). A held-out test set was used to evaluate the effectiveness of the classifiers according to precision, recall and F-measure. In addition, detailed feature analysis was performed to reveal the characteristics of a successful cancer classification model.

*Corresponding author. UQ Health Sciences Building, Royal Brisbane Hospital, Herston, Queensland 4029, Australia. Tel: +61 7 3253 3600.

*Email addresses:* bevan.koopman@csiro.au (Bevan Koopman), g.zuccon@qut.edu.au (Guido Zuccon), Anthony.Nguyen@csiro.au (Anthony Nguyen), anton.bergheim@cancerinstitute.org.au (Anton Bergheim), Narelle.Grayson@cancerinstitute.org.au (Narelle Grayson)

**Results:** The system was highly effective at identifying cancer as the underlying cause of death (F-measure 0.94). The system was also effective at determining the type of cancer for common cancers (F-measure 0.7). Rare cancers, for which there was little training data, were difficult to classify accurately (F-measure 0.12). Factors influencing performance were the amount of training data and certain ambiguous cancers (e.g., those in the stomach region). The feature analysis revealed a combination of features were important for cancer type classification, with SNOMED CT concept and oncology specific morphology features proving the most valuable.

**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.

*Keywords:* Cancer classification, Death certificates, Machine Learning, Natural Language Processing

---

## 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, some of which will describe the type of cancer contributing to death. German et al. [1] demonstrated the importance of analysing death certificates to record cancer-related causes of death for population-based cancer mortality statistics; such statistics from Cancer Registries are vital to measure the effectiveness of healthcare systems and guide cancer control strategies [2]. However, Cancer Registries receive an overwhelming number of death certifi-

cates (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.

In this paper, we propose a system for the automatic classification of cancers from free-text death certificates. The systems has two main components: i) a natural language processing (NLP) pipeline that extracts detailed features (e.g., terms, n-grams, SNOMED CT codes and ICD-O properties) from death certificates; and ii) a set of machine learning classifiers that exploit these features to determine the presence of cancers. The classifiers are deployed in a two-level, cascaded architecture: the first level identified the presence of cancer (i.e., binary cancer/no cancer) and the second level identifies the type of cancer (according to the ICD-10 classification system).

A detailed empirical evaluation on 10 years of semi-manually coded death certificates shows that the proposed system is highly accurate at detecting mentions of cancers (0.942 F-measure on binary classification). The systems is also effective at determining the type of common cancers (average F-measure of 0.7 for the top 20 cancer, which account for 85% of all cancer cases). For rarer cancers, where little data is available to train the classification model, the system is less effective (F-measure of 0.12 on the 15% rarer cancer cases). Finally, a detailed analysis reveals the characteristics of a successful cancer classification model, including the effect of the amount of training data, ambiguity of the terms expressing the cancer, whether the cancer was actually the underlying cause of death and, importantly, the discriminative power of different feature types. The findings of this study helps guide the development of other text classification tasks beyond cancer classification and could be applied to other data sources besides death certificates.

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

## 2. Task Description — Identifying Cancer from Death Certificates

The use case or task proposed in this study has two parts. Give a free-text death certificate, the aims are to 1) determine if cancer was the cause of death; and 2) if it was, determine the type of cancer (according to ICD-10 classification system). Before detailing in the next section 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 recommended by the World Health Organisation [4] and therefore affects how any automated classification is both developed and evaluated. Figure 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, C), often 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 each entry, a duration between onset of the condition and death is stated. For the use case of cancer classification proposed in this study, the sample certificate presented here should firstly be classified as a cancer related death and secondly classified as of type C16 (*Malignant neoplasm of stomach*).[2]

### *2.2. Collection of Death Certificates*

The Cancer Institute of NSW supplied free-text, de-identified death certificates for the years 1999-2008 (inclusive).[3] The certificates were divided into

[2]Note that for this certificate the *underlying* cause of death is taken from Section (I-B), not from the final entry in Section (I-C).

[3]The NSW Population & Health Services Research Ethics Committee granted ethics under application HREC/11/CIPHS/60.

| | |
|---|---|
| (I) | A) HYPOXIC BRAIN INJURY, 5 MINUTES |
| | B) GASTRIC CARCINOMA WITH GASTRECTOMY, 2 MONTHS |
| | C) ATRIAL FIBRILLATION, 6 MONTHS |
| (II) | HYPERCALCAEMIA, 5 YEARS |

Figure 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.

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

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

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 testings 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*

The Australian Bureau of Statistics is responsible for maintaining statistics on causes of death in Australia. This is done by the semi-automatic assignment of ICD-10 codes to death certificates [4]. These ICD-10 codes constitute the ground truth against which the automated classification method is evaluated. For each death certificate, the single underlying cause of death is recorded against that certificate (additional causes of death were not available for this study). All ICD-10 codes were truncated at the three character 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 cases were identified as those certificates assigned any ICD-10 code from ICD-10 Chapter II (*Neoplasms*) [5], including in-situ and benign cancers (i.e., all codes in the range"C00" to "D49"). For individual ICD-10 classification, only "C" codes were considered as these were considered notifiable by the Cancer Registry in which this study was conducted. (A list of these codes and their descriptions is provided in Appendix A.) The frequency distribution according to the type of cancer is shown in Figure 2. The figure shows that a small subset of cancer types make up the vast majority of cancer-caused deaths. In fact, the top 20 most prevalent cancers constitute approximately 85% of all cancer deaths. These top 20 cancers are therefore important to accurately classify with any automated method.

## 3. Feature Extraction Methods

The first component of the automated cancer classification system is a natural language processing pipeline that extracts from a death certificate an array of different features that can be used to train a classification model. A number of different feature types are used; these 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 standard medical terminologies (e.g., the SNOMED CT ontology). The process of extracting concepts from free-text is performed by Medtex, a clinical natural language processing system [6, 7]. Table 2 describes the different types of features extracted, belonging to these two categories. For each feature type, the table columns provide i) a description the feature type; ii) a sample fragment of a death certificate; iii) the resulting features that are consequently derived from the fragment of the death certificate. The feature types listed here were chosen because they were shown to be successful in a previous study on binary (cancer/nocancer) classification of death certificates [8].

Once all features are extracted, death certificates are transformed from orig-

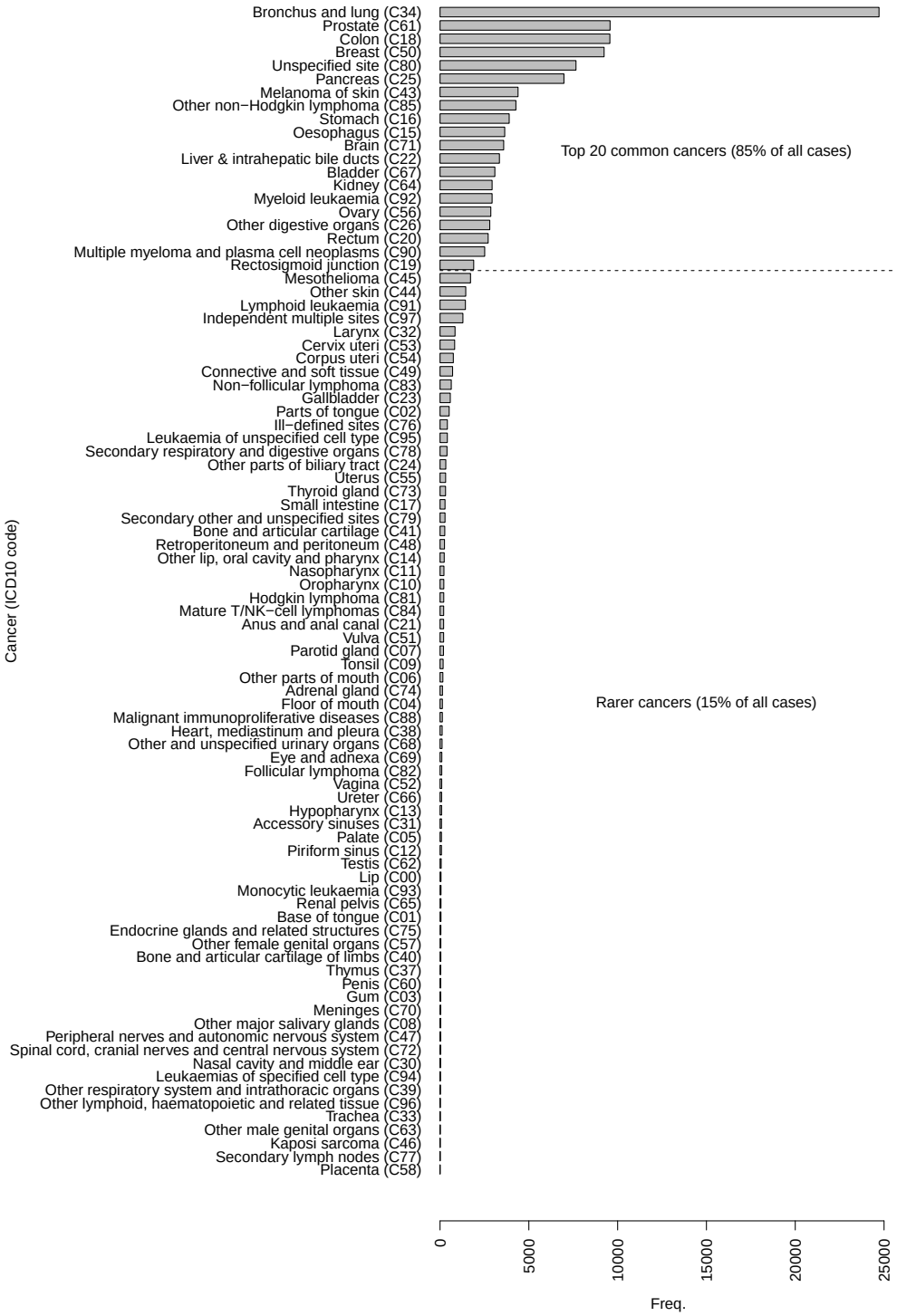





Figure 2: Prevalence of different cancers from ground truth for full set of death certificates (1999–2008).

inal terms to vectors of features (one vector per certificate); 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 if that feature is present in the particular death certificate. Once each death certificate is represented as a feature vector, this feature vector is used as the input to the machine learning classifier.

## 4. Classification Methods

The overall task description proposed in this study is to first identify if cancer was a cause of death and if so determine the type of cancer. We translate this into a machine learning strategy of 1) a single binary classifier that is trained to assign a cancer/nocancer label to a death certificate; and 2) multiple ICD-10 binary classifiers, one for each type of cancer, trained to assign the particular ICD-10 label to a death certificate. The classifiers are deployed in a two-level, cascaded architecture: a death certificate is first issued to the binary cancer/no cancer classifier and if positive the certificate is issued to all the individual ICD-10 classifiers. Thus the binary cancer/nocancer classifier can be referred to as the binary *filter* classifier.

For the implementation of the classifiers we use Support Vector Machines (SVMs).[4] SVMs were chosen as they were the best performing classification model in a previous death certificate classification task [8]. The Weka toolkit was used for the SVM implementation [10]. The parameters for all classifiers were set to the defaults described in Witten et al. [10].

In total, 86 SVMs were developed (85 for ICD-10 and 1 for cancer/nocancer). The cancer/nocancer classifier was trained using the *full* training set; i.e., all the death certificates in the training set (of which 29% were cancers and 71% were

[4] A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

| | 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** | ICDOMorphBerg | Standard ICD Morphology classification system (as defined by [9]) | `METASTATIC ADENOCARCINOMA OF THE LUNG` | 3. Adenocarcinomas |
| | ICDOSiteGroup | Course grained body site descriptions (as defined by Cancer Institute of NSW) | cancer of `lip`, `tongue`, `mouth`, `tonsil` | cancer of *head* region |
| | SCTConceptId | SNOMED CT concept identifier (as extracted by the Medtex system) | `chronic renal failure` | 90688005 |

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.)

non-cancers). In contrast, the ICD-10 classifiers were trained using a *balanced* training set, constructed using the following method: for each ICD-10 code, e.g., CXX, take all the positive CXX cases in the original training set (1999–2006) and include them in a CXX training set with an equal number of non-CXX cancer cases, randomly sampled from the original training set. This was done to ensure that classifiers with small numbers of cases were trained with sufficient attention to positive cases and not skewed by overwhelming number to negative cases [11]. In addition, the use of a balanced approach reduced the computational cost of training, making a large scale evaluation feasible. For comparison to the balanced method, two ICD-10 classifiers (C50 and C34) were also trained using the *full* method. The performance of these (F-measure) was compariable with that of the classifiers trained using the balanced method.

On completion of training, all SVMs were used to classify each certificate in the test set, with the cancer/nocancer classifier used as a filter to the ICD-10 classifiers (i.e., ICD-10 classification were only recorded if there was a preceding positive cancer classification).

## 5. Empirical Evaluation

### *5.1. Evaluation Measures*

Two evaluation measures are considered: precision and recall. Precision (also called positive predictive value) is the fraction of positively classified certificates that are cancer[5], while recall is the fraction of actual cancer certificates that are positively classified.[6] 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 mentioning cancer, while a high recall indicates the system does not miss certificates that contain cancers (particularly important for rare cancers). To provide a single, overall evaluation measure, precision and recall

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

| Precision | Recall | F-measure |
|---|---|---|
| 0.913 | 0.972 | 0.942 |

Table 3: Binary classification results — Evaluation metrics.

| | | Classifier | |
|---|---|---|---|
| | | - | + |
| **Ground truth** | - | 61,278 | 2,622 |
| | + | 789 | 27,482 |

Table 4: Binary classification results — Confusion matrix; + denotes *cancer* and - denotes *nocancer*.

are combined into a third evaluation measure, F-measure.[7]

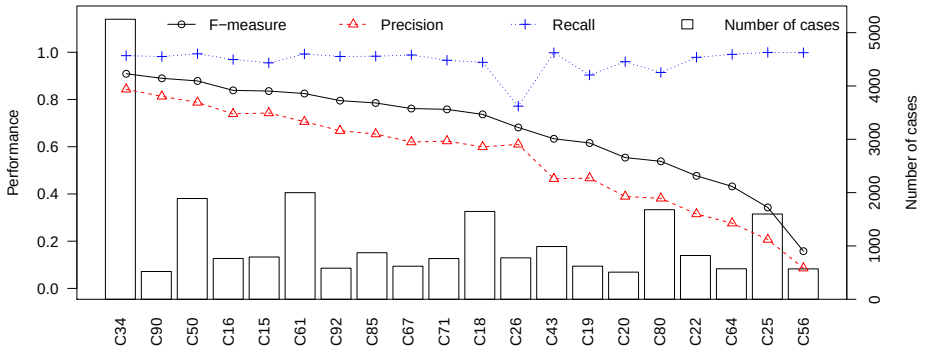
For analysis and interpretation of the ICD-10 classification results, results were divided into two sets, constituting *common* and rare *cancers*. The set of common cancers was derived by: i) ranking ICD-10 classes in descending order of prevalence (according to the ground truth of the testset); and ii) selecting the top $k$ cancers such that 85% of all cancer cases 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 cases.

### *5.2. Classification Results*

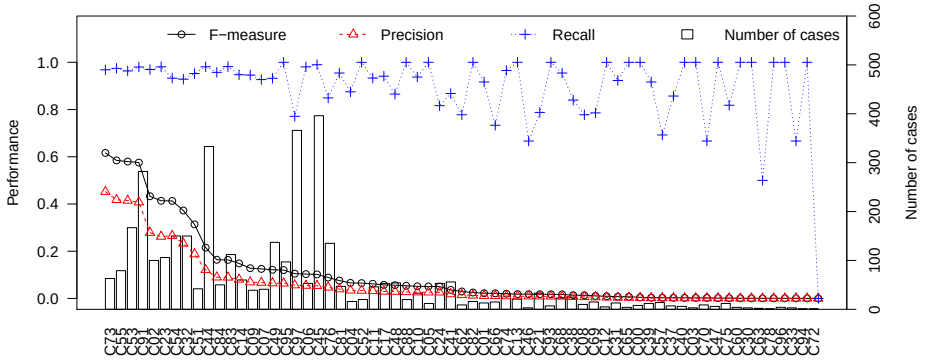
The classification results for the binary (cancer / nocancer) classifier are shown in Table 3. In addition, a confusion matrix, which provides a breakdown of true positives, false positives, true negatives and false negatives, is shown in Table 4. Both precision and recall are high. Precision is reduced by a number of false positives (2,622 in total) and recall reduced by a smaller number of false negatives (789 in total).

The individual ICD-10 classification results are shown in Figure 3 (divided between common and rare cancers). Generally, recall is high but overall effectiveness (F-measure) is reduced by lower precision due to false positives. Performance is better on the common cancers and worse on the rare cancers.

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




(a) Common cancers, constituting 85% of cancer cases.




(b) Rare cancers, constituting 15% of cancer cases.

Figure 3: ICD-10 Classification results: precision, recall and F-measure. Cancers are ordered in descending F-measure.

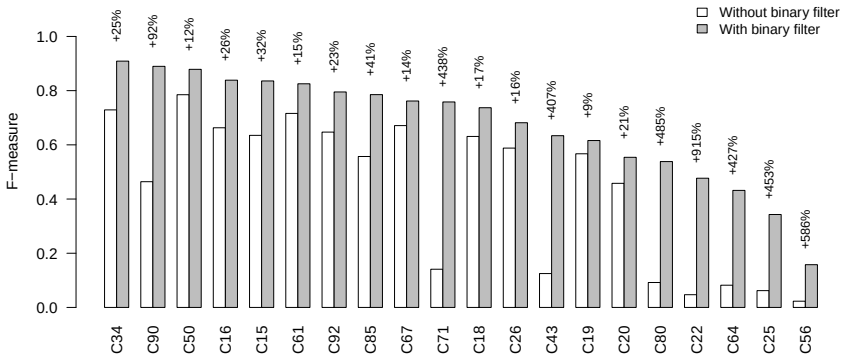



Figure 4: F-measure classification results before and after applying the binary, cancer/nocancer filter. Percentages represent change in F-measure after applying the binary filter.

### *5.3. Effect of Binary Cancer/NoCancer Filter*

The two-level classification method (cancer/nocancer, then ICD-10 classification) was an intentional design decision to improve the effectiveness of the ICD-10 classifier; the purpose being to reduce the number of false positives provided by ICD-10 classification. To quantify the effect that the binary filter had on performance, Figure 4 reports the results of the ICD-10 classifiers with and without the binary filter.[8] The binary filter clearly led to a substantial improvement in classification effectiveness, mainly by reducing false positives in the ICD-10 classifiers (i.e., improvements in precision rather than recall).

## 6. Analysis and Discussion

### *6.1. Classifier Characteristics Impacting Performance*

The results showed that performance was superior on common cancers than on rare cancers. One explanation for this is that the more training data, the better the classifier performance. Figure 5 shows the number of training instances vs. F-measure for each ICD-10 classifier. There was a correlation between training size and performance (Pearson's correlation coefficient was 0.65). However,

[8] For brevity, we only report results for common cancer, although the same trend applied to rare cancers.

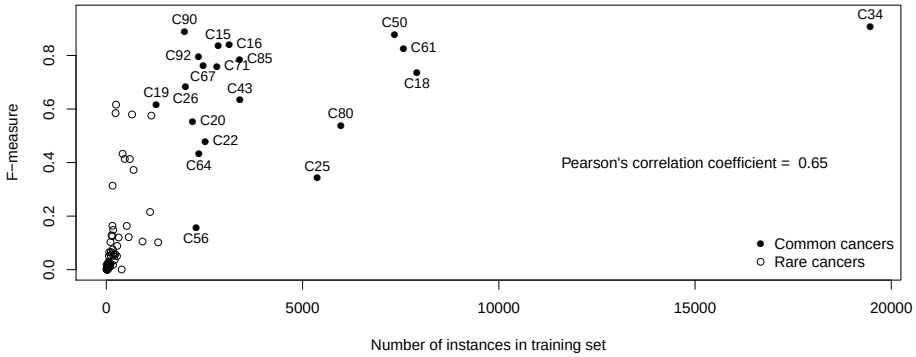



Figure 5: F-measure compared with training set size; correlation 0.65.

this was not the case for all cancers; for example, cancers such as C80 and C25 had a large training set but low F-measure. This shows that some cancers were harder to classify than others (e.g., C25 was *Malignant neoplasm of pancreas* and cancers related to the stomach (spleen, pancreas, liver) were often hard to differentiate). While training data was a determining factor, it was not the only factor. The next section investigates the different feature types which were more discriminative.

### *6.2. Feature Type Analysis*

To understand the value of each of the different feature types (previously outlined in Table 2), we performed a feature analysis study for each of ICD-10 classifiers. Training separate SVMs for each combination of feature type was not feasible (5! feature type combinations $\times$ 85 ICD-10 classes = 10,200 SVMs). Therefore an alternative Information Gain analysis, often applied in machine learning method [10, Sec. 7.1], was applied. Using this method, the worth of a feature was determined by measuring the Information Gain with respect to the predicting class; it can be estimated as $IG(C, F) = H(C) - H(C|F)$, where $C$ is the class (ICD-10 in this case), $F$ is the particular feature of interest and $H$ is the information entropy. The Information Gain of each feature for each ICD-10 class was calculated and features were then ranked by Information Gain, i.e., in descending order of discriminative power. Finally, features were mapped

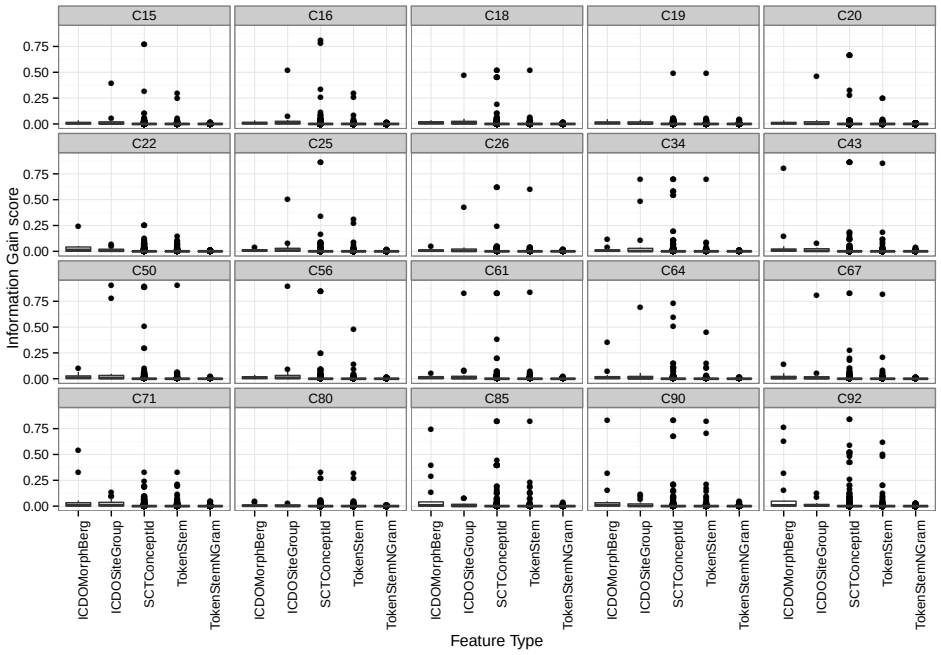
| Rank | Feature | Feature Type | Info. Gain |
|---|---|---|---|
| 1 | `colorect` | TokenStem | 0.423 |
| 2 | `264267007` (Colorectal) | SCTConceptId | 0.422 |
| 3 | `77879006` (Metastatic to) | SCTConceptId | 0.096 |
| 4 | 5.Unspecified carcinomas | ICDOMorphBerg | 0.089 |
| 5 | `79282002` (Carcinoma, metastatic) | SCTConceptId | 0.081 |
| 6 | `metastat` | TokenStem | 0.069 |
| 7 | 17.Unspecified type of cancer | ICDOMorphBerg | 0.066 |
| 8 | `metastat_cancer_colorect` | TokenStemNGram | 0.057 |
| 9 | `14799000` (Neoplasm, metastatic) | SCTConceptId | 0.056 |
| 10 | `cancer` | TokenStem | 0.055 |
| ... | ... | ... | ... |

Table 5: Top 10 features ranked using Information Gain for C19 (*Malignant neoplasm of rectosigmoid junction*) classifier. The most discriminative feature for determining if the cause of death was C19 was the presence of the TokenStem `colorect`, then the presence of the SNOMED CT concept `264267007`.

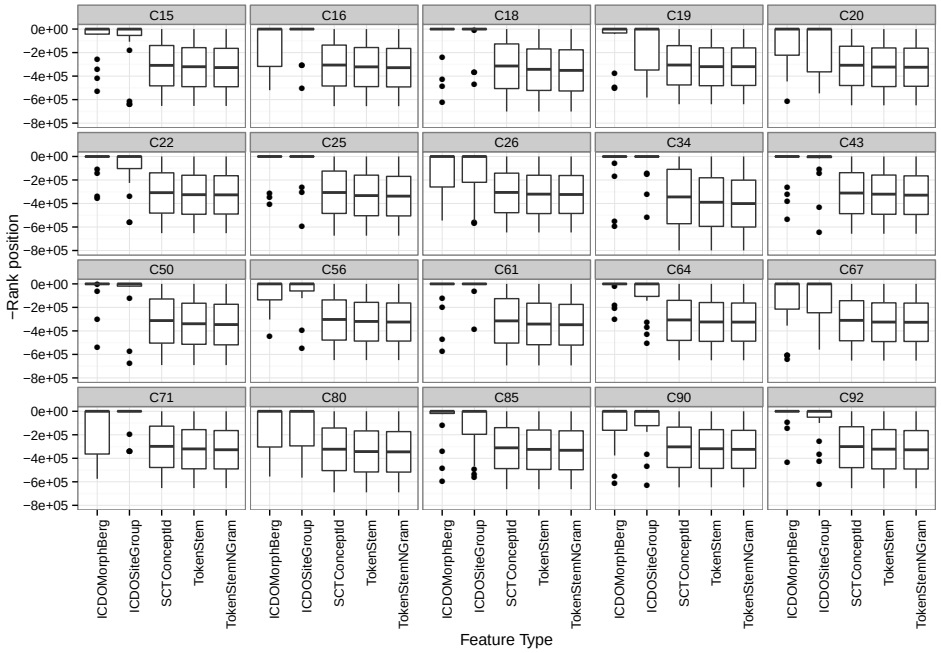
to their respective feature types (e.g., TokenStem, SCTConcept, etc.) in order to determine which feature type was most discriminative. An example of the top 10 features, ranked by Information Gain, for the C19 classifier is shown in Table 5. Using the method described above, both the *rank* and the *score* of different feature types were analysed to determine how effective different feature types were in discriminating each ICD-10 classification.[9]

The different feature types according to Information Gain *score* are shown in Figure 6(a). We observe that there were a number of outliers with a high score from the SCTConceptId feature type. This means that in certain cases (e.g., C16 or C56) the SNOMED CT concept id was a key discriminative feature that indicated this type of cancer; in some cases, this occured together with a high Information Gain for the TokenStem (e.g., C61, C90), while in others it was the SNOMED CT concept alone that was discriminative (e.g., C25). While TokenStem was discriminative in a number of cases (e.g., C18, C34), TokenStemNGram did not prove a strong discriminative feature. Finally, the two cancer specific features, ICDOMorphBerg and ICDOSiteGroup sometimes proved

[9]Note, we are concerned with the rank of the feature *type* not the feature itself; i.e., the "Feature Type" column in Table 5.




(a) Boxplot showing feature type vs. Information Gain score.




(b) Feature type vs. rank position.

Figure 6: Feature type analysis showing the discriminative power of each feature type, ordered by negative rank (-Rank).

valuable (e.g., C56, C71, C90).

The different feature types according to Information Gain negative *rank* are shown in Figure 6(b).[10] Here a different story emerges: features of type ICDOMorphBerg and ICDOSiteGroup were consistently the most discriminative (i.e., were the top-ranked features). The other three feature types, SCTConceptId, TokenStem and TokenStemNGram, were consistently found lower in the ranking of discriminative features, with SCTConceptId slightly above TokenStem and TokenStem slightly above TokenStemNGram.

The difference between the score results of Figure 6(a) and the negative rank results of Figure 6(b) revealed that although SCTConceptId and TokenStem had some high Information Gain scores at the top of ranking, they also had many low scores at the bottom of the ranking. This means they were not consistently reliable features. In addition, the ICDOMorphBerg and ICDOSiteGroup were far less common (i.e., they only occur for some death certificates) but that when they did occur they were important indicators for the correct cancer-related cause of death for that certificate.

In summary, no single feature type was discriminative across the board for ICD-10 classification. The cancer specific ICDOMorphBerg and ICDOSiteGroup were particularly discriminative but cannot be relied upon as they did not occur in every certificate. The SCTConceptId feature generally proved more valuable than TokenStem; TokenStemNGram added little.

## 7. Limitations and Future Work

The ground truth ICD-10 code for a death certificate (described in Section 2.3) was the single underlying cause of death for that certificate; ICD-10 codes were not available for additional causes of deaths listed in the certificate. For some certificates, cancer may have been recorded in the text of the death certificate but may not have been the underlying cause of death and there-

[10] Negative rank position was used so that the most discriminative features (i.e., those with low rank numbers) appear at the top of the plot.

| | |
|---|---|
| (I) | A) MYOCARDIAL INFARCTION, MINUTES |
| | B) DIABETES, 6 YEARS |
| (II) | BREAST CANCER, 6 YEARS |

Figure 7: Death certificate listing breast cancer as a contributing cause of death but with the primary cause of death being the diabetes listed in Section (I)(B).

fore the ground truth would not record this certificate as being cancer related. For example, consider the death certificate presented in Figure 7. In this case, based on the presence of breast cancer in the text, the machine learning classifiers would have marked the certificate as being cancer, thus leading to a false positive classification. A consequence of this would be that the evaluation underestimated the effectiveness of the classifiers (certainly, false positives were the main source of errors). Two avenues of future work can be pursued to address this issue. Firstly, alternative ground truth that records both underlying and additional causes of death may be sort out, however such information is currently not available and may be laborious to develop from scratch. Secondly, and alternative to new ground truth, the feature extraction and classification methods in this study can be extended to take into account the death certificate structure (i.e., Section (I) or (II)), using this as an additional feature in the classification task. (Note that the underlying cause is not always simply the last entry in Section (I); instead, there are a series complex rules that are employed when determining the underlying cause of death [12].)

The evaluation was done independently for each ICD-10 classification. In reality, the system would run in parallel with each classifier assigning a positive or negative label to a death certificate. In this case, it is possible for a single certificate to be assigned multiple ICD-10 codes. If the ground truth contains a single underlying cause of death then a method is required to *fuse* the multiple ICD-10 classifications results into a final ICD-10 code. One fusion method is to use the SVM distance from the hyperplane to assign a probability of correctness and select the classifier with highest likelihood. Other fusion methods may bias the choice based on the type of cancer (e.g., to ensure rare cancers are identified

by the system). The investigation of fusion methods is currently being pursued as future work.

The empirical results showed poor performance on very rare cancers where very little training data was available. For these cases a supervised machine learning approach may not be desirable. Symbolic, rule-based approaches have been successful in other cancer identification tasks [13] and may be applied here. Using the fusion method approach described in the previous paragraph, a hybrid approach, rules in conjunction with the SVMs, could be used; the fusion method could use evidence from both the rules and the SVMs in determining the final classification for a certificate.

Finally, the feature analysis based on Information Gain revealed key discriminative features. Many features had an Information Gain score of zero (i.e., they provided no indication as to what classification label to apply to a certificate). Future work will investigate a feature reduction method based on Information Gain to considerably reduce the feature space and produce more efficient SVMs (both in terms of training time and memory usage: time and space). This is an important requirement for Cancer Registries who have to efficiently process large volume of death certificates.

### *7.1. 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 [14]; in Australia, cancer notifications and synoptic reporting are performed automatically from pathology and cytology reports [6]. These case studies show there is both a need and viable use case for automated classification of cancers from cancer registry data.

There have been a number of text mining applications specifically focusing on extracting cancer related information; Spasic et al. [15] provides a comprehensive review of these. The review highlighted a strong bias towards symbolic

techniques, i.e., the use of pattern matching and dictionary lookup for cancer-related entity extraction. A number of symbolic approaches make us of some natural language processing and pattern matching aided by some medical domain knowledge resource, either the UMLS Metathesaurus [16, 17] or some other resource [18]. The survey of these approaches identified some limiting factors for symbolic techniques: i) the effort required to develop manual rules, which have to be defined on a case-by-case basis for each cancer type; and ii) brittleness of symbolic approach to the idiosyncrasies of the clinical sublanguage such as non-standard abbreviations as well as a high degree of spelling and grammatical errors. The authors conclusion is that a shift from rule-based methods to machine learning is required.

Some machine learning methods have been applied to text classification for cancer. Butt et al. [8] developed a binary (i.e., cancer or no-cancer) classifier for free-text death certificates. They investigated a number of different statistical classifiers. These classifiers were evaluated on a small 5,000 report subset of data from Cancer Institute New South Wales: this data is similar to that used in the present article. They found that Support Vector Machines performed best in determining whether cancer was the underlying cause of death: this provided initial evidence for using Support Vector Machine classifiers in our work. There are three important distinctions between this previous study and the work described in this paper: i) we developed a set of ICD-10 classifiers that determined the type of cancer, whereas Butt et al. simply developed a binary classifier for the presence of cancer (not the type); ii) we conducted an extensive empirical investigation on a large dataset to determine the robustness and general applicability of our methods; and iii) we analysed the factors and features that affected the performance of automated cancer classification.

## 8. Conclusion

This study provides a system for automatically identifying and characterising cancers 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 proposed system has 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 is very effective in determining if cancer was a cause of death (F-measure of 0.942 on binary cancer/nocancer) and is also effective at determining the type of common cancers (average F-measure of 0.7), while rare cancers are more challenging (F-measure of 0.12). The two-level architecture of first identifying if cancer is present and then determining the type of cancer (according to ICD-10) is an important method to improve the accuracy of the ICD-10 classification.

Although the amount of training data was a factor influencing performance (performance was worse for rarer cancers), it was not the only influencing factor and that certain cancers (e.g., those in the stomach region) were harder to classify than others. Detailed feature analysis via Information Gain revealed that no single feature type is most discriminative in determining the type of cancer, although the use of SNOMED CT concepts and ICD-O morphology and site features proved most valuable. For rarer cancers, a symbolic rule-based approach may be more suited; this method can be included with SVMs in a hybrid method that *fuses* results from different SVMs and rules to determine a final ICD-10 classification. This is an active area of future work.

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

**Authors' contributions**

AB, AN 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 performed the empirical evaluation, analysis of results. BK and GZ drafted

the manuscript. All authors reviewed and approved the final manuscript.

### Statement on 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, D. Yin, The accuracy of cancer mortality statistics based on death certificates in the united states, Cancer epidemiology 35 (2) (2011) 126–131.

[2] M. Coleman, D. Forman, H. Bryant, J. Butler, B. Rachet, C. Maringe, U. Nur, E. Tracey, M. Coory, J. Hatcher, 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, The Lancet 377 (9760) (2011) 127–138.

[3] Cancer Council Australia, Cancer in Australia: Facts and Figures (2014) [cited 14 October 2014].
URL `http://www.cancer.org.au/about-cancer/what-is-cancer/facts-and-figures.html`

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

[5] World Health Organization, International statistical classification of diseases and related health problems, Vol. 1, World Health Organization, 2004.

[6] A. Nguyen, J. Moore, M. Lawley, D. Hansen, S. Colquist, Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications., in: Health Informatics Conference, Vol. 168, Brisbane, 2011, pp. 117–124.

[7] 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.

[8] L. Butt, G. Zuccon, A. Nguyen, A. Bergheim, N. Grayson, Classification of cancer-related death certificates using machine learning, The Australasian medical journal 6 (5) (2013) 292.

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

[10] I. H. Witten, E. Frank, Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, 2005.

[11] K. E. Chai, S. Anthony, E. Coiera, F. Magrabi, Using statistical text classification to identify health information technology incidents, Journal of the American Medical Informatics Association (2013) amiajnl–2012.

[12] Centers for Disease Control and Prevention, Instructions for Classifying the Underlying Cause-of-Death, ICD-10 (2014).

[13] A. N. Nguyen, M. J. Lawley, D. P. Hansen, R. V. Bowman, B. E. Clarke, E. E. Duhig, S. Colquist, Symbolic rule-based classification of lung cancer stages from free-text pathology reports, Journal of the American Medical Informatics Association 17 (4) (2010) 440–445.

[14] M. L. Gjerstorff, The danish cancer registry, Scandinavian journal of public health 39 (7 suppl) (2011) 42–45.

[15] I. Spasić, J. Livsey, J. A. Keane, G. Nenadić, Text mining of cancer-related information: Review of current status and future directions, International journal of medical informatics 83 (9) (2014) 605–623.

[16] 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.

[17] K. Davis, C. Staes, J. Duncan, S. Igo, J. C. Facelli, Identification of pneumonia and influenza deaths using the death certificate pipeline, BMC medical informatics and decision making 12 (1) (2012) 37.

[18] 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 medical informatics and decision making 12 (1) (2012) 88.

## Appendix A. ICD-10 Code Descriptions

ICD-10 codes covering cancer considered in this study.

| | |
|---|---|
| C00 | Malignant neoplasm of lip |
| C01 | Malignant neoplasm of base of tongue |
| C02 | Malignant neoplasm of other and unspecified parts of tongue |
| C03 | Malignant neoplasm of gum |
| C04 | Malignant neoplasm of floor of mouth |
| C05 | Malignant neoplasm of palate |
| C06 | Malignant neoplasm of other and unspecified parts of mouth |
| C07 | Malignant neoplasm of parotid gland |
| C08 | Malignant neoplasm of other and unspecified major salivary glands |
| C09 | Malignant neoplasm of tonsil |
| C10 | Malignant neoplasm of oropharynx |
| C11 | Malignant neoplasm of nasopharynx |
| C12 | Malignant neoplasm of piriform sinus |
| C13 | Malignant neoplasm of hypopharynx |
| C14 | Malignant neoplasm of other and ill-defined sites in the lip, oral cavity and pharynx |
| C15 | Malignant neoplasm of oesophagus |
| C16 | Malignant neoplasm of stomach |
| C17 | Malignant neoplasm of small intestine |
| C18 | Malignant neoplasm of colon |
| C19 | Malignant neoplasm of rectosigmoid junction |
| C20 | Malignant neoplasm of rectum |
| C21 | Malignant neoplasm of anus and anal canal |
| C22 | Malignant neoplasm of liver and intrahepatic bile ducts |
| C23 | Malignant neoplasm of gallbladder |
| C24 | Malignant neoplasm of other and unspecified parts of biliary tract |
| C25 | Malignant neoplasm of pancreas |
| C26 | Malignant neoplasm of other and ill-defined digestive organs |
| C30 | Malignant neoplasm of nasal cavity and middle ear |
| C31 | Malignant neoplasm of accessory sinuses |
| C32 | Malignant neoplasm of larynx |
| C33 | Malignant neoplasm of trachea |
| C34 | Malignant neoplasm of bronchus and lung |
| C37 | Malignant neoplasm of thymus |
| C38 | Malignant neoplasm of heart, mediastinum and pleura |
| C39 | Malignant neoplasm of other and ill-defined sites in the respiratory system and intrathoracic organs |
| C40 | Malignant neoplasm of bone and articular cartilage of limbs |
| C41 | Malignant neoplasm of bone and articular cartilage of other and unspecified sites |
| C43 | Malignant melanoma of skin |
| C44 | Other malignant neoplasms of skin |
| C45 | Mesothelioma |
| C46 | Kaposi sarcoma |

| | |
|---|---|
| C47 | Malignant neoplasm of peripheral nerves and autonomic nervous system |
| C48 | Malignant neoplasm of retroperitoneum and peritoneum |
| C49 | Malignant neoplasm of other connective and soft tissue |
| C50 | Malignant neoplasm of breast |
| C51 | Malignant neoplasm of vulva |
| C52 | Malignant neoplasm of vagina |
| C53 | Malignant neoplasm of cervix uteri |
| C54 | Malignant neoplasm of corpus uteri |
| C55 | Malignant neoplasm of uterus, part unspecified |
| C56 | Malignant neoplasm of ovary |
| C57 | Malignant neoplasm of other and unspecified female genital organs |
| C58 | Malignant neoplasm of placenta |
| C60 | Malignant neoplasm of penis |
| C61 | Malignant neoplasm of prostate |
| C62 | Malignant neoplasm of testis |
| C63 | Malignant neoplasm of other and unspecified male genital organs |
| C64 | Malignant neoplasm of kidney, except renal pelvis |
| C65 | Malignant neoplasm of renal pelvis |
| C66 | Malignant neoplasm of ureter |
| C67 | Malignant neoplasm of bladder |
| C68 | Malignant neoplasm of other and unspecified urinary organs |
| C69 | Malignant neoplasm of eye and adnexa |
| C70 | Malignant neoplasm of meninges |
| C71 | Malignant neoplasm of brain |
| C72 | Malignant neoplasm of spinal cord, cranial nerves and other parts of central nervous system |
| C73 | Malignant neoplasm of thyroid gland |
| C74 | Malignant neoplasm of adrenal gland |
| C75 | Malignant neoplasm of other endocrine glands and related structures |
| C76 | Malignant neoplasm of other and ill-defined sites |
| C77 | Secondary and unspecified malignant neoplasm of lymph nodes |
| C78 | Secondary malignant neoplasm of respiratory and digestive organs |
| C79 | Secondary malignant neoplasm of other and unspecified sites |
| C80 | Malignant neoplasm, without specification of site |
| C81 | Hodgkin lymphoma |
| C82 | Follicular lymphoma |
| C83 | Non-follicular lymphoma |
| C84 | Mature T/NK-cell lymphomas |
| C85 | Other and unspecified types of non-Hodgkin lymphoma |
| C88 | Malignant immunoproliferative diseases |
| C90 | Multiple myeloma and malignant plasma cell neoplasms |
| C91 | Lymphoid leukaemia |
| C92 | Myeloid leukaemia |
| C93 | Monocytic leukaemia |
| C94 | Other leukaemias of specified cell type |
| C95 | Leukaemia of unspecified cell type |
| C96 | Other and unspecified malignant neoplasms of lymphoid, haematopoietic and related tissue |
| C97 | Malignant neoplasms of independent (primary) multiple sites |