Health Informatics: Building a Healthcare Future Through Trusted Information
A.J. Maeder and F.J. Martin-Sanchez (Eds.)
IOS Press, 2012

doi: 10.3233/978-1-61499-078-9-150

Classification of pathology reports for cancer registry notifications

Anthony NGUYEN [a,1], Julie MOORE [b], Guido ZUCCON [a], Michael LAWLEY [a], and Shoni COLQUIST [b]
[a] *The Australian E-Health Research Centre, CSIRO ICT Centre, Brisbane, Australia*
[b] *Queensland Cancer Control Analysis Team, Queensland Health, Brisbane, Australia*

Abstract. Objective: To develop a system for the automatic classification of pathology reports for Cancer Registry notifications. **Method:** A two pass approach is proposed to classify whether pathology reports are cancer notifiable or not. The first pass queries pathology HL7 messages for known report types that are received by the Queensland Cancer Registry (QCR), while the second pass aims to analyse the free text reports and identify those that are cancer notifiable. Cancer Registry business rules, natural language processing and symbolic reasoning using the SNOMED CT ontology were adopted in the system. **Results:** The system was developed on a corpus of 500 histology and cytology reports (with 47% notifiable reports) and evaluated on an independent set of 479 reports (with 52% notifiable reports). Results show that the system can reliably classify cancer notifiable reports with a sensitivity, specificity, and positive predicted value (PPV) of 0.99, 0.95, and 0.95, respectively for the development set, and 0.98, 0.96, and 0.96 for the evaluation set. High sensitivity can be achieved at a slight expense in specificity and PPV. **Conclusion:** The system demonstrates how medical free-text processing enables the classification of cancer notifiable pathology reports with high reliability for potential use by Cancer Registries and pathology laboratories.

Keywords. automatic data processing, data mining, disease notification, natural language processing, neoplasm, systematized nomenclature of medicine

Introduction

Cancer is a notifiable disease in Queensland and other States and Territories in Australia. The Queensland Cancer Registry (QCR) is a population based register of diagnosed cancer cases in Queensland from 1982. All public and private pathology laboratories are legally required under the Public Health Act 2005 to provide copies of specimen reports that contain a result of cancer to the Queensland Cancer Registry. Traditionally, this process involves the manual identification of notifiable cancer reports by the pathology laboratory, which are mailed to the Cancer Registry for cancer notification processing. Despite the fact that electronic pathology reporting is the norm at pathology laboratories, the notification of reports to the QCR is a manual and paper based process. Pathology laboratories notify the majority of cancer notifications and provide the most reliable and detailed basis for a diagnosis of cancer.

---

[1] Corresponding Author: Anthony Nguyen, PhD, The Australian e-Health Research Centre, Level 5 – UQ Health Sciences Building 901/16, Royal Brisbane and Women's Hospital, Herston QLD 4029, Australia; E-mail: anthony.nguyen@csiro.au

With the updated technology changes in clinical information management, sending and receiving electronic pathology HL7 feeds from pathology laboratories is now available in Queensland Health. Given this technological advancement, there is a need to filter the pathology feed into reports that are cancer notifiable and those that are not. Reconciling the pathology feed with the incoming notifications from pathology laboratories by means of a report identifier is a viable solution. Alternatively, an automated computer-assisted approach could be used to automatically identify pathology reports that are notifiable. The latter approach potentially provides significant benefits to both the Cancer Registry and the individual pathology laboratories.

This paper investigates an automated process for filtering notifiable pathology reports from non-notifiable ones. Pathology HL7 message filtering, Queensland Cancer Registry business rules, natural language processing, and symbolic reasoning using SNOMED CT[2] subsumption querying were incorporated to achieve high sensitivity (98%) and specificity (95%) on a nearly balanced dataset of 979 pathology reports.

1. Background

Critical to the cancer notification process is the ability to identify notifiable pathology reports. Pathology laboratories are required to notify a specimen from all positive histology (including haematology) and cytology reports, excluding urine, sputum and pap smears. In Queensland, notifiable cancers to the registry include:

1. All invasive cancers excluding basal cell carcinoma (BCC) and squamous cell carcinomas (SCC) of the skin;
2. Any cancer with uncertain behavior;
3. All in-situ conditions; and
4. Benign central nervous system and brain tumours.

The notifications received by the Cancer Registry would subsequently be abstracted for cancer specific information such as primary site, histological type and grade, etc. This information is used to form an automated consolidation of a patient cancer record in the Registry. The notification of cancer reports also extend to supporting documentation, herein called supporting notifiable reports, such as follow-up pathology reports that include wider excisions or lymph node removal resulting in no residual cancer or those that have equivocal results.

1.1. Automatic Filtering of Notifiable Pathology Reports

A number of systems have been proposed to address the automatic filtering of notifiable pathology reports. *E-path* is a cancer finding and reporting system that selects those that contain reportable cancer findings and determines applicable codes for each report [1]. Although, conceptually similar to the proposed approach, it is unclear how much adaptation is required to support Australian Cancer Registries.

The Case Finding Engine (CaFE) [2] scans reports for custom made list of cancer related terms, phrases and SNOMED codes. Phrases indicating negative findings were also used to rule out cancer cases. The sensitivity of the system was reported to be 1.00,

[2] Systematized nomenclature of medicine - clinical terms

while specificity was 0.85. The custom list of terms was institution specific and would need to be adapted to capture the wording and spelling variations in other institutions.

Open Registry [3] is a system that selects reports with disease codes indicating cancer. This approach assumes that all reports are coded (and reliable), which is not usually the case as observed in the pathology data in this study.

The Automated Retrieval Console (ARC) [4] uses supervised machine learning algorithms to classify pathology or radiology reports that were "consistent with cancer". Good performances ranging from an F-measure of 0.75 for lung cancer to 0.94 for colorectal cancer were achieved without custom code or rules development. As the machine learning algorithms were domain specific, a large number of classifiers would potentially need to be developed.

MEDTEX is a innovative system initially designed for the classification of QLD Cancer Registry notifications data from pathology reports [5]. High reliability in the classification of notifiable reports were achieved with sensitivity, specificity, and positive predicted value of 0.97, 0.99, and 0.99, respectively for the development set, and 1.00, 0.96, and 0.87 for the evaluation set. The analysis on the classification of notifiable reports, however, was based on a limited dataset and was not the main thrust of the study.

This research extends the promising MEDTEX system and evaluates the task of notifiable report classification on a larger dataset. Unlike previous approaches, the system does not rely solely on custom phrases, explicitly mentioned disease codes, or the development of tumour specific classification models.

2. Method

2.1. System Architecture

Figure 1 shows the high-level architecture of the proposed cancer notifiable pathology report classification system. A two pass approach is proposed: The first pass queries for pathology report types that are required by the QLD Cancer Registry, while the second pass analyses the free text in reports to identify those reports that are cancer notifiable.

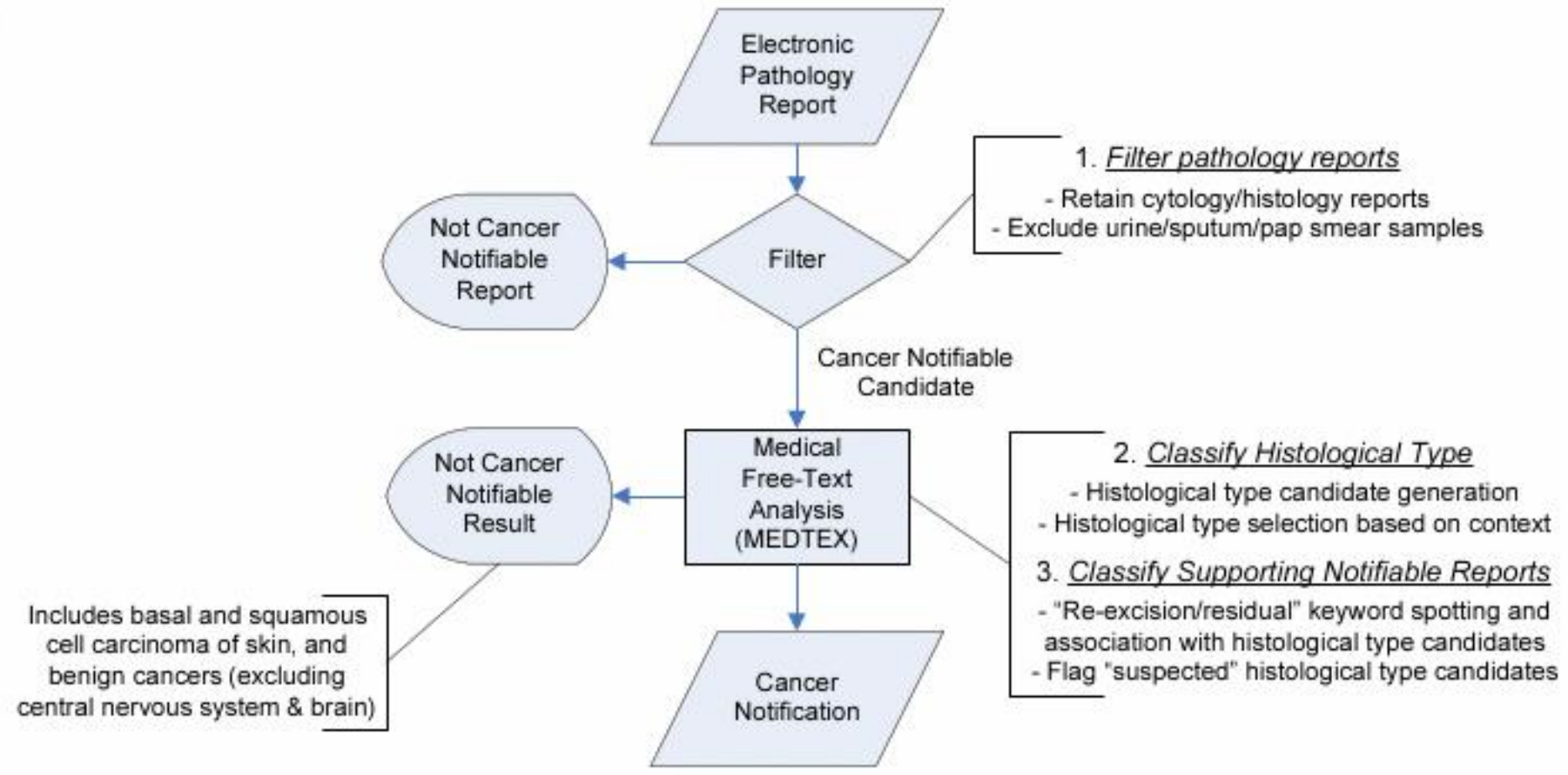

Figure 1. Proposed system high-level architecture.

2.1.1. Filtering of Pathology Reports

The filtering of pathology reports retrieves report types that are potentially notifiable (including supporting notifiable reports) to the QCR; these include histology (and haematology) and cytology reports (excluding urine, sputum and pap smears). Table 1 lists the reports that are potentially notifiable. A query was executed to achieve this.

Table 1. Pathology tests or observations relating to histology or cytology reporting.

Report Type	Pathology Tests or Observation	
Histology	Bone marrow	Histology Frozen
	BM Asp & Treph	Histology Biopsy
Cytology	Cytology (Skin,D/C)	Cytology FNA
	Cytology (Fluids)	Flow Cytometry

2.1.2. Classifying Notifiable Cancers

The Medical Text Extraction (MEDTEX) system [5, 6] was used to identify cancers with histological types that were notifiable to the Cancer Registry. A two pass approach was proposed. The first pass iterates through the SNOMED CT concepts identified in the free text from the non-history sections of the report, and reasons over them to see if they are subsumed by (or a descendent of) the top level concepts identified to be notifiable to the Cancer Registry (see Table 2). Legacy SNOMED ID codes found in the free text often refers to the cancer's histological type; however, these concepts are ignored as these codes are not mandatory and often incomplete.

Table 2. Selected top-level SNOMED CT concepts for histological type candidate generation.

Concept ID	Fully Specified Name
367651003	Malignant neoplasm of primary, secondary, or uncertain origin (morphologic abnormality)
127569003	In situ neoplasm (morphologic abnormality)
86251006	Neoplasm, uncertain whether benign or malignant (morphologic abnormality)
253061008	Nervous system tumor morphology (morphologic abnormality)
128928004	Neuroendocrine neoplasm (morphologic abnormality)
115241005	Neuroepitheliomatous neoplasm (morphologic abnormality)

In the second pass, histological type candidates were associated with contextual phrases surrounding the concept to assert the concept as either "present", "absent", or "possible". For those candidates that have been asserted as "present", the most advanced histological type (based on the concept's ICD-O morphology code) was selected as the most likely histological type for the report. These reports would be classified as notifiable; unless the histological type was found to be SCC or BCC and associated with a "skin" concept (i.e. concepts co-occurred within the same sentence).

2.1.3. Classifying Supporting Notifiable Reports

Supporting notifiable reports include those with excisions resulting in no residual cancer, re-excisions and suspected notifiable cancers. If no histological type was found (and the report was not a BCC or SCC of skin), then the histological type candidates found would either be an empty set, or asserted as either "absent" or "possible". Using these findings, any of the following conditions would trigger the classification of a report as a supporting notifiable report:

1. At least one histological type candidate was asserted as "possible";

2. At least one histological type candidate was associated (i.e. co-occurred within the same sentence) with the keyword "residual"; or
3. Keyword "re-excision" was found in the report.

2.2. Corpus Description

A corpus of 979 histology and cytology reports was obtained with research ethics approval from the Queensland Health Research Ethics Committee. Each report satisfied the initial filtering pass in Section 2.1.1

The reports covered a large range of cancers with approximately 50% of the reports being notifiable. The breakdown of the corpus into a development and evaluation set is shown in Table 3, along with the breakdown of the notifiable reports into whether it contained a notifiable cancer or supporting notifiable information. The ground truth was created based on an adjudication process between the reference data set provided by a domain expert and the output of the system for all reports in the development and evaluation set.

Table 3. Corpus statistics.

	Development Set	Evaluation Set
Notifiable (Canc./Supp.)	237 (201/36)	248 (220/28)
Non-Notifiable	263	231
	500	479

Canc., Reports containing notifiable cancers;
Supp., Reports containing supporting notifiable information

3. Results

Table 4 shows the overall performance of the automated notifiable classification system in terms of sensitivity (or recall), specificity, positive predicted value (PPV; or precision) and F-measure (balanced F-score or F1 score). Here, sensitivity and specificity are statistical measures of how well the system correctly identifies positive (notifiable) and negative (non-notifiable) cases, respectively. While PPV is a measure of the proportion of the system predicted notifiable cases that are in fact notifiable in the ground truth, and F-measure (F) is interpreted as a weighted average of precision (P) and recall (R) given by $F = 2PR/(P+R)$.

The misclassifications from the system is shown in Table 5 as a confusion matrix where the frequency counts according to assigned ("System") class labels and actual ("Ground Truth") class labels are tabulated. From the development set, 16 reports were misclassified (error rate = 3.2%), while 14 were misclassified in the evaluation set (error rate = 2.9%).

Table 4. Notifiable classification performances.

	Sensitivity	Specificity	Positive Predictive Value	F-measure
Development Set	0.9873	0.9506	0.9474	0.9669
Evaluation Set	0.9839	0.9567	0.9606	0.9721

Table 5. Confusion matrices comparing notifiable report classifications.

Development Set	System	
	Notif.	Not Notif.
Ground Truth		
Notif.	234	3
Not Notif.	13	250

Notif., notifiable

Evaluation Set	System	
	Notif.	Not Notif.
Ground Truth		
Notif.	244	4
Not Notif.	10	221

Notif., notifiable

4. Discussion

Overall, the automatic cancer notifiable classification system performed reliably with sensitivity, specificity, PPV and an F-measure of 0.987, 0.951, 0.947 and 0.967 respectively for the development set, and 0.984, 0.957, 0.961 and 0.972 for the evaluation set. Good performances in both the development and evaluation set (with a low error rate) show that the approach is sufficient and robust.

The confusion matrices show that a small number of false negatives (missed notifications; 7 cases 0.72%) were classified at an expense of a moderate number of false positives report notifications (23 cases; 2.35%) from both the development and evaluation set. Although there were limited missed notifications, the cost of these would likely be much higher than false positive notifications. Error analysis reveals that of the 7 missed notifications, 5 were potentially diagnostically relevant to the notification of cancers while 2 were from supporting notifiable reports. Of the 5 that were potentially diagnostically relevant, 2 were incorrectly identified as SCC or BCC of skin, 2 were due to the incorrect application of negation phrases to the histological type concepts, and 1 was due to an "immunohistochemistry" supplementary report where the report substructure confused the system in thinking that the histological type concept was part of the history section. Further gain in sensitivity could potentially be achieved by addressing the above negation and report substructure issues. The BCC and SCC of skin rule could also be removed from the system to improve sensitivity; however this would be at the expense of increased false notifications.

In terms of the 23 false notifications, error analysis reveal that 10 were due to misclassifications as supporting notifiable reports, 3 were meant to be BCC or SCC of skin, while the remaining false notifications were incorrectly classified as notifiable due to other issues relating to the classification algorithm.

5. Conclusion

The proposed system demonstrates how medical free-text processing could enable the classification of electronic cancer notifiable pathology reports with high reliability for potential use by Cancer Registries and pathology laboratories. Queensland Cancer Registry business rules, natural language processing, and symbolic reasoning over the text using the SNOMED CT ontology were adopted in the system. Results show that the approach is promising for the cancer notifiable classification of pathology reports.

References

[1] D. Dale, *et al.*, "The impact of E-path technology on Ontario Cancer Registry operations," *J Registry Manag.*, 29:52-56, 2002.
[2] D. Hanauer, *et al.*, "The Registry Case Finding Engine (CaFE): An automated approach for cancer patient identification from unstructured, free-text pathology reports.," *J Clin Oncol,* 24:320s-320s, 2006.
[3] P. Contiero, *et al.*, "Comparison with manual registration reveals satisfactory completeness and efficiency of a computerized cancer registration system," *J Biomed Inform,* 41:24-32, 2008.
[4] L. D'Avolio, *et al.*, "Evaluation of a Generalizable Approach to Clinical Information Retrieval using the Automated Retrieval Console (ARC)," *J. Am. Med. Inform. Assoc.,* 17:375-382, 2010.
[5] A. Nguyen, *et al.*, "Automatic Extraction of Cancer Characteristics from Free-Text Pathology Reports for Cancer Notifications," in *Health Informatics Conference,* 2011, pp. 117-124.
[6] A. Nguyen, *et al.*, "Symbolic rule-based classification of lung cancer stages from free-text pathology reports," *J. Am. Med. Inform. Assoc.,* 17:440-445, 2010.