Health Informatics: The Transformative Power of Innovation
D.P. Hansen et al. (Eds.)
IOS Press, 2011

doi: 10.3233/978-1-60750-791-8-117

Automatic Extraction of Cancer Characteristics from Free-Text Pathology Reports for Cancer Notifications

Anthony NGUYEN[a, 1], Julie MOORE[b], Michael LAWLEY[a], David HANSEN[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 Cancer Registry notifications data from free-text pathology reports. *Method*: The underlying technology used for the extraction of cancer notification items is based on the symbolic rule-based classification methodology, whereby formal semantics are used to reason with the systematised nomenclature of medicine – clinical terms (SNOMED CT) concepts identified in the free text. Business rules for cancer notifications used by Cancer Registry coding staff were also incorporated with the aim to mimic Cancer Registry processes. *Results*: The system was developed on a corpus of 239 histology and cytology reports (with 60% notifiable reports), and then evaluated on an independent set of 300 reports (with 20% notifiable reports). Results show that the system can reliably classify notifiable reports with 96% and 100% specificity, and achieve an overall accuracy of 82% and 74% for classifying notification items from notifiable reports at a unit record level from the development and evaluation set, respectively. *Conclusion*: Cancer Registries collect a multitude of data that requires manual review, slowing down the flow of information. Extracting and providing an automatically coded cancer pathology notification for review can lessen the reliance on expert clinical staff, improving the efficiency and availability of cancer information.

Keywords. Automatic Data Processing, Data Mining, Disease Notification, Neoplasm, Systematised Nomenclature of Medicine

Introduction

Cancer is a notifiable disease in all States and Territories in Australia and is the only major disease category from which an almost complete coverage of incidence data is available. Cancer notification is an important and fundamental tool for providing an accurate picture of the impact of cancer, and the nature and extent of cancer. The Cancer Registry data is used to help improve cancer prevention and control to improve treatments and survival rates for patients with 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

In response to a manual and paper based Queensland Cancer Registry (QCR) cancer notification process and updated technology changes in clinical information management, an automated computer system to simplify the cancer notifications process was developed. We have focused on a computer-aided cancer notification tool that incorporates the extraction and coding of cancer notification information embedded within pathology reports by identification and tracking of patient cancer cases from electronic pathology feeds in a state-wide oncology information repository.

1. Background

The traditional method of cancer notification and minimum data set abstraction by Cancer Registry personnel has been to manually scan and read paper copies of pathology cancer specimen reports. An extensive amount of scientifically validated clinical information is required to be abstracted and coded for each report.

The information is often trapped in unstructured, ungrammatical, and often fragmented free-text. It has been found that there is repeated interpretation of the results with differing conclusions. The effort required for reading and interpreting the results is extremely labour and time intensive and subject to errors of omission.

By employing a semi-automated system, the reliance on expert clinical coders can be lessened, thus improving the efficiency and availability of health information. Improving the current cancer notifications process would provide significant benefits to oncology service providers, health administrators, researchers and patients.

1.1. Automating Cancer Notifications

Critical to the collation of cancer notifications data is the ability to identify cancer related pathology reports, extract information from the reports relating to the cancer notification's minimum data set, provide cancer notifications if a new cancer incident was found, and track the patient's progress as more pathology reports are received for the same person.

A number of systems have been developed to address various aspects of the cancer notification's workflow, namely:

- E-path, a commercially available computerised cancer finding and electronic pathology reporting product. It consists of mapping pathology data to HL7 messages, determining applicable codes for each report and scanning reports to select those that contain reportable cancer findings [1]. E-path is only applicable to parts of the cancer notifications business process in Queensland and does not satisfy the beginning to end processes. The other issue with commercial products is the potential large initial cost and ongoing maintenance that is involved in designing and implementing solutions for particular business requirements.
- Case Finding Engine (CaFE) [2], a tool to automatically scan free text medical documents for custom made list of terms, phrases and SNOMED codes relevant to cancer. Phrases indicating negative findings were also used to rule out cancer cases. Relevant terms were highlighted for review by a registrar and used to populate the registry database. For pathology reports, the sensitivity for automated case identification was 100%, but specificity was 85.0%. However,

the custom made list of terms requires customisation to capture wording and spelling variations unique to each institution wanting to implement the system.

- Open Registry [3] software generates incidence data by scanning hospital discharges, pathology reports, and death certificates, and selecting those with disease codes indicating cancer. Cancer registrations were either accepted automatically or flagged for manual review with the aim to speed up data production and enhance registry efficiency. This approach assumes that all pathology reports are coded (and are reliable), which is not usually the case as observed in the pathology data used in this study.

1.2. Clinical Language Processing of Pathology Reports

Related studies in extracting key cancer characteristics from pathology free text include the medical text analysis system/pathology (MedTAS/P) proposed by Coden et al [4]. Natural language processing (NLP), machine learning and rules were used to capture cancer disease characteristics and populate a cancer disease knowledge representation model (CDKRM). Selected cancer characteristics were evaluated and showed promise when evaluated against a corpus of pathology reports for patients with colon cancer.

Within the Cancer Biomedical Informatics Grid initiative, the Cancer Tissue Information Extraction System (caTIES) [5] aims at extracting concepts from surgical pathology reports using the National Cancer Institute Enterprise Vocabulary System and MetaMap [6]. caTIES was developed for the purpose of indexing and retrieval.

Another system designed to improve the processes of information retrieval is the Automated Retrieval Console (ARC) [7]. ARC processes and converts the unstructured text to structured data such as SNOMED or UMLS codes, and subsequently used as features for supervised machine learning. The study showed that good performance can be achieved for the classification of pathology or radiology reports that are "consistent with cancer" in the domains of colorectal, prostate and lung.

In previous work, we have also reported on the use of machine learning and symbolic rule based approaches for the extraction of information from free-text pathology reports, specifically for lung cancer synoptic reporting and staging [8-10]. The next section presents the symbolic rule-based approach in more detail and applies it to the task of cancer notifications for all cancer types.

2. Method

2.1. System Description

The system automatically scans HL7 messages from the Queensland Oncology Repository (QOR). It identifies the type of report (histology or cytology) and looks for cancer characteristics such as primary site and laterality, histological type and grade, histology collection date and basis of diagnosis that is relevant to the notification of cancers. Non-notifiable cancers such as non-malignant cancers, squamous cell carcinoma (SCC), and basal cell carcinoma's (BCC) of the skin were also identified, but removed and flagged by the system.

The proposed system builds upon the Medical Text Extraction (MEDTEX) system [11], a software platform for the development of language engineering analysis engines for natural language applications, such as information extraction. MEDTEX incorporates domain knowledge to bridge the gap between natural language and the use of clinical terminology semantics for automatic medical text inference and reasoning. These services turn the medical narrative into structured data which can be easily stored, queried or rendered by most systems for use in their health application. Analysis engines using the MEDTEX technology have been developed to:

- standardise the free text by identifying medical (or more specially SNOMED CT [12]) concepts in the free text, including abbreviations and acronyms, shorthand terms, dimensions and relevant legacy codes,
- relate key medical concepts, terms and codes using contextual information and report substructure such as the detection and application of medical negation phrases to relevant disease and finding concepts, and the segmentation of the free text into sections, and
- use formal semantics to reason with the clinical concepts; inferring complex clinical notions relevant to a health application such as the extraction of cancer characteristics important for cancer notifications.

The semantic inference and reasoning techniques developed exploit the report substructure and make use of the semantics encoded in SNOMED CT concepts. This forms the foundation for the semantic analysis of concepts identified in the free text. Subsumption relationships in SNOMED CT [13] can be taken advantage of to identify concepts which relate to the same disease, anatomy or finding, or infer if two descriptions relate to the same concept. The underlying technology used for the extraction of cancer notification items is based on the symbolic rule-based classification methodology presented in [8, 10], which has been successfully applied to lung cancer synoptic reporting and staging from pathology reports.

Business rules for cancer notifications used by Cancer Registry Coders [14] were also incorporated with the aim to mimic the processes adopted by the QLD Cancer Registry. Cancer notification data element values were classified by using ICD-O Third Edition [15] for primary site, histological type and grade, and other data elements according to classification codes recorded in the QLD Cancer Registry [14].

2.2. Extraction of Notification Items

A two pass approach was used to classify the following cancer notification items:

- Primary site (ICD-O topography code in the form Cxx.x; ranging from C00.0 to C80.9)
- Laterality (right, left, bilateral, not applicable, and unknown; code numbers 1, 2, 3, 8 and 9, respectively)
- Histological type (ICD-O morphology code in the form M-xxxxx; first 4 digits ranging from 8000 to 9989, and the last digit can be 0-3)
- Histologic grade (9 classes; code numbers from 1 to 9)
- Basis of diagnosis (4 classes; code numbers from 06 to 09)

The first pass detects candidate notification items from the report substructure and SNOMED CT concepts identified from the pre-processing steps in MEDTEX. Only concepts relating to the current pathological examination (i.e. non-history sections of the

report) were processed. Legacy SNOMED ID's recorded in the free text for topographies and morphologies were ignored by the system since they were found to be lacking specificity or not recorded.

The second pass then uses the QLD Cancer Registry coding business rules to filter the candidate concepts and to select the most likely concept for the notifications data element. Synoptic headings, if present in the report, such as 'location' or 'site', 'grade', etc are also used to help restrict the scope of search for notification items.

A combination of the following methods was used to classify the cancer notification data elements:

- QLD Cancer Registry coding business rules (including special casings),
- domain knowledge assumptions,
- SNOMED CT properties to expose morphology codes,
- SNOMED CT to ICD-O topography code cross-maps,
- concept subsumption queries against relevant top-level concepts,
- querying concept attributes for relevant concepts such as procedure and finding sites,
- ad-hoc association queries to search the SNOMED CT ontology, and
- keyword/phrase spotting.

The methods involving the manipulation of SNOMED CT concepts form the basis of the symbolic rule-based classification technique [8].

Histological type was first classified and used to filter non-notifiable reports. It was also used to restrict the scope of search at a sentence level for other notification items. If no relevant notifications items were classified within the restricted scope, then appropriate sections (i.e. microscopic or macroscopic, then summary) were used to extend the scope of search.

The gold standard used for system evaluation was based on an adjudication (or error analysis) between the reference data set (or silver standard) provided by a clinical coder and the output of the system. The error analysis is used to better understand the business rules governing the cancer notification process, bounds of abstraction accuracies by humans, and feedback the type of errors generated by the system for further development.

3. Results

A baseline system was developed on an initial hand-selected corpus of 239 histology and cytology reports. The reports covered a broad range of cancers (including rare ones and no cancers), where 60% of the reports were notifiable cancers. The evaluation set, on the other hand, contained a random selection of 300 histology & cytology reports, where only 20% of the reports were notifiable. The reports were obtained from a state-wide pathology information system within Queensland Health with research ethics approval from the Queensland Health Research Ethics Committee.

The baseline system shows high reliability in the classification of notifiable reports with 96% and 100% specificity from the development and evaluation set, respectively. All non-notifiable reports were assigned an "NA" to all notification fields. Only 1 report (0.4%) from the development and 9 reports (3%) from the evaluation set were misclassified by the system as notifiable, while only 4 reports (1.7%) from the

development set were misclassified by the system as non-notifiable. The system did not make any misclassifications on the notifiable reports in the evaluation set.

The adjudication process corrected 49 and 32 labels from the development and evaluation set, respectively (i.e. 3% error rate from the clinical coder). Note that no reports were available in either the development or evaluation data set for testing the basis of diagnosis of "autopsy and histology" (code = 09).

Table 1 presents the accuracy measures for only the notifiable reports from the gold standard. In the set of notifiable reports, 39 and 22 labels from the development and evaluation sets respectively were corrected during the adjudication process, again approximately, a 3% error rate from the clinical coder.

Table 1. Results for key cancer notification items for notifiable reports only.

Cancer Notification Item	Development Set (N = 140) Accuracy % (95% CI*)	Evaluation Set (N = 61) Accuracy % (95% CI)
Basis of Diagnosis	96 (90–98)	92 (81–97)
Histological Type	81 (73–87)	66 (52–77)
Histological Grade	94 (89–97)	87 (75–94)
Primary Site	60 (51–68)	51 (38–64)
Laterality	76 (68–82)	75 (62–85)
Summary	**82** (75–88)	**74** (61–84)

*All 95% confidence intervals are calculated using the Wilson procedure.

For Cancer Registry reporting of cancer incidence, the primary site is reported at a site (Cxx; rather than at a sub-site Cxx.x) level (e.g. "C50 Breast"). In addition, a number of sites may be grouped for reporting purposes such as "C33, C34 Trachea, Lung" and "C19, 20 Rectosigmoid junction and Rectum." Histological types are also used to group the sites for cancer incidence reporting (e.g. "M-959, M-967 to M-972 Non Hodgkin Lymphoma"). Table 2 presents the systems results when the "reportable" primary sites are used. These results, when compared to Table 1, show the proportion of cases where the system would classify a similar histology or site to the ground truth, but was only incorrect in specifying the histology's exact characteristics or primary site's sub-site

Table 2. Results for key cancer notification items from notifiable reports only, based on the cancer registry groupings of primary site for cancer incidence reporting.

Cancer Notification Item	Development Set (N = 140) Accuracy % (95% CI)	Evaluation Set (N = 61) Accuracy % (95% CI)
Basis of Diagnosis	96 (90–98)	92 (81–97)
Histological Type	81 (73–87)	66 (52–77)
Histological Grade	94 (89–97)	87 (75–94)
"Reportable" Primary Site	*75 (67–82)*	*80 (68–89)*
Laterality	76 (68–82)	75 (62–85)
Summary	**85** (78–90)	**80** (68–89)

4. Discussion

Due to the large difference in the proportions of notifiable reports between the two data sets, only the notifiable reports were evaluated. The system achieved an overall accuracy of 82% and 74% on the development and evaluation sets, respectively. Identifying histological type and primary site was a challenge for the system, given the limited number of development reports and the large number of possible histological types and primary sites.

Although, the histological type and primary site algorithms require further development, the system's output for these notification items were observed in many cases to be similar in histology or site to the ground truth. By grouping primary sites at a site and histology level, similar to that used for Cancer Registry reporting of incidence, the performance of the system is enhanced significantly to 85% and 80% for the development and evaluation set, respectively.

Future research will need to focus on improving the system's performance, especially for critical notification items such as primary site and histological type. Additional hand-selected reports covering known "problem" cases will need to be used to further iteratively develop the system. The system's coverage of coding rules as per QLD Cancer Registry clinical coding manual will also need increased coverage. In addition, the system could be extended to handle multiple primary sites per notifiable report; the current data set is limited in such cases and did not provide sufficient examples for its development.

The generation of patient level notifications by combining report level notifications have been implemented but its evaluation and comparison with Cancer Registry notifications is beyond scope of this paper. It is hoped that through larger scale evaluations such as the patient level evaluations, other insights could be gained to further refine the system.

Other considerations include the practical significance of the system where there is a need for the automated "triage" of reports, where classification confidence scores would need to be assigned to the system's results for expert review. There is also a need for high (or 100%) specificity in the classification of notifiable reports. Although the system achieved 100% specificity in the classification of notifiable reports in its evaluation set, the development set contained a number of cases that the system misclassified as non-notifiable.

5. Conclusion

The cancer notifications system is currently under development and preliminary results for the extraction and coding of notifiable items show promise. Formal semantics are used to reason with the SNOMED CT concepts identified in the free text. Business rules for cancer notifications used by QLD Cancer Registry coding staff were also incorporated with the aim to mimic Cancer Registry processes. The system is conjectured to support the clinical coding workflow at Cancer Registries by highlighting and pre-populating cancer notification items for review. It is hoped that enabling improved decision support, will assist by improving efficiency, reducing costs arising from duplicated processes and timeliness of cancer information.

References

[1] D. Dale, et al., "The impact of {E-path} technology on {Ontario Cancer Registry} operations", *Journal of Registry Management*, vol. 29, pp. 52–56, 2002.

[2] D. A. Hanauer, et al., "The Registry Case Finding Engine (CaFE): An automated approach for cancer patient identification from unstructured, free-text pathology reports", *Journal of Clinical Oncology*, vol. 24, pp. 320s–320s, Jun 20 2006.

[3] P. Contiero, et al., "Comparison with manual registration reveals satisfactory completeness and efficiency of a computerised cancer registration system", *J Biomed Inform*, vol. 41, pp. 24–32, Feb 2008.

[4] A. Coden, et al., "Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model", *Journal of Biomedical Informatics*, vol. 42, pp. 937–949, Oct 2009.

[5] R. S. Crowley, et al., "caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research", *J Am Med Inform Assoc*, vol. 17, pp. 253–64, May 1 2010.

[6] A. R. Aronson and F. M. Lang, "An overview of MetaMap: historical perspective and recent advances", *J Am Med Inform Assoc*, vol. 17, pp. 229–36, May 1 2010.

[7] L. D'Avolio, et al., "Evaluation of a Generalisable Approach to Clinical Information Retrieval using the Automated Retrieval Console (ARC)", *Journal of the American Medical Informatics Association*, vol. 17, pp. 375–382, 2010.

[8] A. N. Nguyen, et al., "Structured pathology reporting for cancer from free text: lung cancer case study", in *18th Annual Health Informatics Conference, HIC 2010*, Melbourne, Australia, 2010, pp. 69–72.

[9] I. A. McCowan, et al., "Collection of cancer stage data by classifying free-text medical reports", *Journal of the American Medical Informatics Association*, vol. 14, pp. 736–745, Nov-Dec 2007.

[10] A. N. Nguyen, et al., "Symbolic rule-based classification of lung cancer stages from free-text pathology reports", *Journal of the American Medical Informatics Association*, vol. 17:, pp. 440–445, July 1, 2010; 2010.

[11] A. N. Nguyen, et al., "A Simple Pipeline Application for Identifying and Negating SNOMED Clinical Terminology in Free Text", in *17th Annual Health Informatics Conference, HIC 2009*, Canberra, Australia, 2009, pp. 188–193.

[12] International Health Terminology Standards Development Organisation, "SNOMED Clinical Terms User Guide", 2008.

[13] International Health Terminology Standards Development Organisation, "SNOMED Clinical Terms – Transforming Expressions to Normal Forms", Jan 31 2007.

[14] Queensland Cancer Registry, "Clinical Coding Manual Version 3".

[15] World Health Organisation, *International Classification of Diseases for Oncology*, 3 ed. Geneva: World Health Organisation, 2007.