Automated Cancer Registry Notifications: Validation of a Medical Text Analytics System for Identifying Patients with Cancer from a State-Wide Pathology Repository

Anthony N. Nguyen, PhD[1], Julie Moore[2], John O'Dwyer[1], Shoni Philpot[2]
[1]The Australian e-Health Research Centre, CSIRO, Brisbane, Australia;
[2] Queensland Cancer Control Analysis Team, Department of Health, Queensland Government, Brisbane, Australia

Abstract

The paper assesses the utility of Medtex on automating Cancer Registry notifications from narrative histology and cytology reports from the Queensland state-wide pathology information system. A corpus of 45.3 million pathology HL7 messages (including 119,581 histology and cytology reports) from a Queensland pathology repository for the year of 2009 was analysed by Medtex for cancer notification. Reports analysed by Medtex were consolidated at a patient level and compared against patients with notifiable cancers from the Queensland Oncology Repository (QOR). A stratified random sample of 1,000 patients was manually reviewed by a cancer clinical coder to analyse agreements and discrepancies. Sensitivity of 96.5% (95% confidence interval: 94.5-97.8%), specificity of 96.5% (95.3-97.4%) and positive predictive value of 83.7% (79.6-86.8%) were achieved for identifying cancer notifiable patients. Medtex achieved high sensitivity and specificity across the breadth of cancers, report types, pathology laboratories and pathologists throughout the State of Queensland. The high sensitivity also resulted in the identification of cancer patients that were not found in the QOR. High sensitivity was at the expense of positive predictive value; however, these cases may be considered as lower priority to Cancer Registries as they can be quickly reviewed. Error analysis revealed that system errors tended to be tumour stream dependent. Medtex is proving to be a promising medical text analytic system. High value cancer information can be generated through intelligent data classification and extraction on large volumes of unstructured pathology reports.

Introduction

The Queensland Cancer Registry (QCR) is a population based cancer registry that monitors and records the incidence and mortality of cancer in the State of Queensland, Australia over time by collecting cancer notifications from a variety of sources. The QCR collection, therefore, has a number of intrinsic values, which it derives from its population base, and provides the capacity to support longitudinal analysis. The QCR also supports key activities in Queensland and nationally such as health service planning and cancer research.

Pathology laboratories throughout Queensland are legally required to notify the Cancer Registry of pathology tests that contain a result of cancer. 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 behaviour;
3. All in-situ conditions; and
4. Benign central nervous system and brain tumours.

All of the pathology cancer notifications are currently paper based and the incidence of cancers is increasing with the number of new cancer cases in Queensland increasing by more than 177% between 1982 and 2012; the growth in new cancer cases is largely due to population growth and ageing[1]. A growing backlog is delaying the delivery of more timely cancer information due to the extent of manual processing and an out-dated information collection system. This highlights the need for introducing new technologies to assist with automating cancer registry processes.

With some updated technology changes, pathology laboratories can now send pathology electronically via HL7 feeds to the Queensland Oncology Repository (QOR). However, this is still not without its challenges. There is no mechanism to determine whether an electronic pathology message is cancer notifiable or not, apart from reading and interpreting the contents of the HL7 pathology report. As such, a computer-assisted approach is required to automatically identify pathology reports that are cancer notifiable.

Background

A number of systems have been proposed to address the automatic detection of cancer notifiable (or reportable) pathology reports. These systems rely solely on the use of custom made list of cancer and non-cancer related terms, phrases, and disease codes that may be institution specific, or the development of reportable cancer or tumour specific machine learning classification models[2-5]. The proposed method however aims to be generic supporting the full range of tumour streams (or types of cancers) from a breadth of pathology laboratories, report types and pathologists.

*C*ommercial cancer finding and reporting systems that selects reports that contain reportable cancer findings based on dictionary and linguistic analysis are also available[6-7]. Although, conceptually similar to the proposed approach, it is unclear as to how much modification is required to achieve high levels of accuracy within the context of an Australian Cancer Registry.

The proposed approach is based on Medtex[8]. Medtex is an emerging medical text analytic capability that conducts intelligent data classification and extraction on large volumes of unstructured pathology reports, which can result in the generation of high value cancer information. Medtex has been developed to automate Cancer Registry notification using QCR business rules, natural language processing, and symbolic reasoning using SNOMED CT subsumption querying[9-13]. Medtex can assist in supporting the continuous improvement of cancer notification and enables improved decision support for Registry clinical coders.

Related work on automating cancer notifications using Medtex has shown that the system classified cancer notifiable reports with a sensitivity, specificity, and positive predictive value (PPV) of 0.98, 0.96, and 0.96, respectively, for an evaluation set of 479 histology and cytology reports[9]. Although very promising results were achieved, outperforming many of the systems in the literature, the dataset was very limited in size. The reports were selectively sampled such that an approximately balanced dataset resulted with half of the dataset covering a range of cancers, while the other half containing non-cancers. In spite of this, Medtex aims to be generic and was developed using Cancer Registry business rules and the normalisation and reasoning of reports using standard clinical terminologies; it does not rely on custom phrases, explicitly mentioned disease codes, or the development of machine-learning classification models[2-5].

This research extends previous work and evaluates the utility of Medtex to identify notifiable cancer pathology reports on a larger and representative cancer population based dataset. Results show that high sensitivity and specificity can be achieved for a representative dataset containing a breadth of tumour streams, report types, pathology laboratories and pathologists. Analysis of discrepancies revealed that Medtex was able to 1) identify cancer patients not found in the QOR potentially underestimating the incidence of cancer and 2) actual system errors tended to be tumour stream dependent. These insights were not possible with the previous study that used a smaller and unrepresentative dataset. These insights also allow for specific system limitations to be addressed in future developments.

Method

System description

A high-level architecture of the Medtex medical text analytic system for cancer notifications classification and data extraction is shown in Figure 1.

In this paper, the cancer notification classification component of the system will be evaluated. A two-pass approach is adopted to classify whether pathology reports were cancer notifiable or not[9]. The first pass queries and filters pathology HL7 messages from the pathology data repository to identify report types that are required by the Cancer Registry, while the second pass analyses the free text reports and classifies them as either "Cancer Notifiable" (i.e., positive histology or cytology reports, excluding urine, sputum and pap smears), "Not Cancer Notifiable" or "Supporting Notifiable Reports" (i.e., those with further excisions resulting in no residual cancer, re-excisions and/or suspected notifiable cancers).

More specifically, Medtex is developed using the General Architecture for Text Engineering (GATE) platform[14]. It unifies the language in pathology reports by mapping spans in the text (e.g., clinical terms, abbreviations and acronyms, short-hand terms and relevant legacy codes) into SNOMED CT concepts[8]. Cancer notification identification is then based on finding the most likely primary site and histological type of the cancer based on SNOMED CT subsumption querying, and analyzing the report substructure and contextual information surrounding the concept to ensure that the concept's mention was not modified by negative or uncertain assertions. Following QCR business rules, histological types that were either squamous or basal cell carcinoma (SCC or BCC) and was associated with the skin were classified as "Not Cancer Notifiable"[16]. Here, if SCC or BCC concepts were associated with a "skin" concept (i.e., concepts co-occurred within the same sentence), then such a classification would be made. The

QCR business rules are consistent and align with those nationally for the notification of cancers. More details on the cancer notification classification component of the system can be found in previous work[9].

The architecture supporting Medtex is characterised by a messaging framework built on the concept of message queues, producers and consumers[11,15]. Because multiple message consumers can be set up in parallel to receive messages from the same queue, Medtex provides high throughput, making it an ideal framework for analysing large streams of electronic pathology feeds.

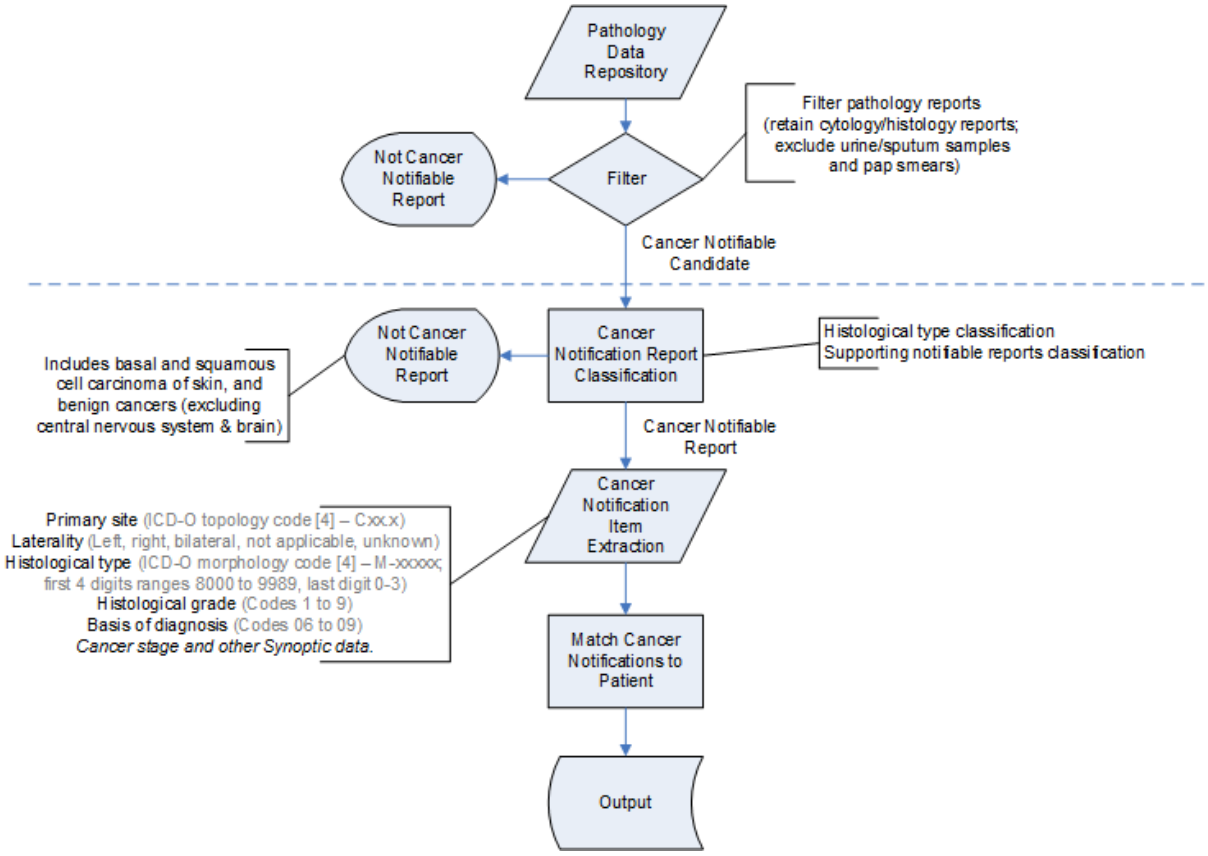

Figure 1. High-level architecture of cancer notification classification and data extraction modules in Medtex.

The producer queries for pathology report types, which are potentially notifiable to the Cancer Registry (i.e., histology and cytology), and adds them to the input queue for Medtex processing. Medtex can then be set up either as a single instance or in parallel as consumers acting on the input queue, where each consumer will take a message from the input queue in turn for processing and analysis. The results from the Medtex analysis are then published to an output queue where another consumer can subscribe to the output queue to store the results in a SQL database. Figure 2 shows the database output schema portraying the range of cancer notification information extracted from pathology reports such as primary site of cancer, histology, grade, laterality, metastatic sites, as well as synoptic data including stage[9-13].

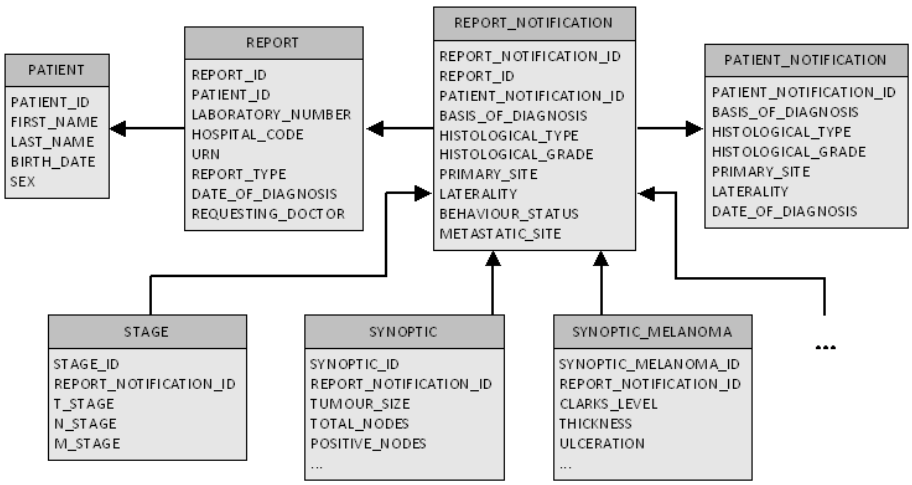

Figure 2. Cancer notifications database output schema portraying the range of cancer notification information extracted from pathology reports.

Data collection and processing

With research ethics approval from the Queensland Health Research Ethics Committee, a corpus of 45.3 million HL7 pathology messages from the Queensland Oncology Repository (QOR), for the year of 2009 was used. QOR compiles and collates data from a range of Queensland cancer data source systems including the Queensland Cancer Registry (QCR), hospital admissions data, death data, treatment systems, public and private pathology and various hospital clinical data systems. The QOR is a resource managed by the Queensland Cancer Control Analysis Team (QCCAT) within the Department of Health, Queensland.

The corpus comprises of pathology laboratory results in the State of Queensland from the public health sector; these include bloods, tumour markers, etc. A total of 100 out of 114 possible unique report codes (or report types) were identified in the corpus. Upon filtering of the HL7 pathology feeds for histology and cytology report types (16 report types in total; e.g., histology frozen, histology biopsy, cytology (fluids), flow cytology, haematology) that are potentially relevant for Cancer Registry notifications, 119,581 histology and cytology reports remained for Medtex processing.

The histology and cytology reports were reported from 34 pathology laboratories in Queensland and therefore cover a breadth of report types, tumour streams (including non-cancer cases), laboratories and pathologists.

Performance evaluation

The 119,581 histology and cytology reports were analysed as a batch process by Medtex for reports requiring notification to the Cancer Registry.

These reports were consolidated at a patient level based on the exact matching of their first and last names, sex and date of birth (DOB). Patients with invalid or potentially multiple first/last name, sex and DOB combinations were removed from the analysis; for example, patients with punctuations in names (e.g., commas, quotes, hyphens), unknown sex and DOB. Patients with a combination of the above, for example, patients having same first name, sex and DOB but a different last name due to a hyphen (e.g., "FirstName LastName" and "FirstName MaidenName-LastName" with same DOB and Sex, and patients having the same first and last name and DOB as those with an unknown sex), were also removed from the analysis. The removal of such patients was to ensure that that the full set of reports processed by Medtex for a patient most likely corresponded to the set of reports for patients used by the Cancer Registry for cancer notifications coding.

A patient was classified as "Cancer Notifiable" if Medtex classified at least one of its reports as "Cancer Notifiable", and vice versa, if Medtex classified all reports for a patient as "Not Cancer Notifiable", then the patient was classified as "Not Cancer Notifiable". For the purposes of the evaluations conducted in this paper, "Supporting Notifiable Report" classifications by Medtex were assumed "Cancer Notifiable" for the patient-level consolidation process. This assumption is likely to over estimate the number of patient-level cancer cases resulting from Medtex.

The QOR was used as the ground truth to identify actual patients diagnosed with cancer. A patient was considered to be "Cancer Notifiable" if the patient existed in the QOR database, and vice versa, the patient was assumed "Not Cancer Notifiable" if the patient could not be found in the QOR.

A cancer clinical coder reviewed a stratified sample of 1,000 patients (each with potentially multiple reports) to analyse the agreements and discrepancies between Medtex and the QOR. A larger number of patients were selected from the stratum where Medtex and the QOR were in disagreement. This was purposely done to study the potential limitations of the system. The cancer clinical coder manually reviewed all the reports pertaining to the patients independently and classified their findings as "Cancer Notifiable", "Not Cancer Notifiable" and "Supporting Notifiable Reports" in a similar way that Medtex classifies them. Again, results from a report-level were consolidated at a patient-level for analysis. Discrepancies between the coder's judgements and that from Medtex and the QOR were analysed to uncover the cause of the discrepancy.

An adjusted contingency table (or confusion matrix) of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) was generated (in light of the review of discrepancies) to compute the sensitivity (or recall), specificity, positive predictive value (or precision) and F1 score of Medtex in classifying cancer notifications.

Results

Medtex filtered the 45.3 million HL7 pathology messages and processed and analysed the 119,581 histology and cytology reports in just under 5 days using 3 Medtex instances running in parallel (average processing rate of 3.6 seconds per report). A range of cancer characteristics were extracted, however, the current evaluation is only concerned with the cancer notification classification output by Medtex.

Upon patient consolidation and the removal of 1,359 patients whose reports could not be reliably aggregated at a patient-level based on the exact match of first and last name, sex and DOB, 85,502 unique patients remained for the evaluation against cancer patients from the QOR.

The contingency table showing the distribution of matches and mismatches between Medtex and the QOR at a patient-level is shown in Table 1.

Table 1. Contingency table of patient matches and mismatches between Medtex and the Queensland Oncology Repository (QOR).

		Medtex		Total
		Patient Notifiable	Patient Not Notifiable	
QOR	Cancer Patient Exists	12,799	11,021	23,820
	Cancer Patient Not Exist	3,178	58,504	61,682
Total		15,977	69,525	85,502

As there were a high number of discrepancies between the information contained in the QOR and the classification output by Medtex, a stratified sample of 1,000 patients was randomly selected from the quadrants of the contingency table for manual review by an experienced cancer clinical coder. The breakdown of the patients and their reports are summarised in Table 2.

Table 2. Breakdown of the number of patients and reports for manual coding.

	# Patients	# Reports
QOR Patient Exist / Medtex Notifiable	150	322
QOR Patient Not Exist / Medtex Notifiable	350	638
QOR Patient Exist / Medtex Not Notifiable	350	469
QOR Patient Not Exist / Medtex Not Notifiable	150	180
Total	1,000	1,609

A summary of the results and analyses are reported in the following section. The aim of the analyses is two-fold: 1) to generate an adjusted contingency table to determine the indicative performance of the system, and 2) identify insights and system limitations for ongoing research and development. The classification results from the cancer clinical coder and Medtex and associated detailed error analyses can be made available upon request.

1. Analysis of Medtex patient notifiable classification and patient exists in the QOR

The manual review of 150 patients that exist in the QOR and were also classified as "Cancer Notifiable" by Medtex revealed that there were 8 patients whose reports did not suggest the need for a cancer notification; these patients existed in the QOR but for cancers diagnosed in other years (not in 2009). An error analysis on these Medtex misclassified cases were observed to fall into two categories, namely, 1) Medtex misclassifying SCC/BCC of the skin, which are to be excluded from Cancer Registry notifications, and 2) other algorithm related issues from Medtex such as the missed detection of negative assertion phrases. Due to the 8 false positive cases, there is an error rate of 8/150 = 0.0533 (95% Confidence Interval (CI)[1]: 0.025-0.1059). Extrapolating these false positive findings within the respective patient stratum revealed that 682 patients (i.e. error rate x 12,799; 95% CI: 320-1,356)) were Medtex notifiable but indeed did not have cancer; thus requiring adjustment.

2. Analysis of Medtex patient not notifiable classification and patient not exist in the QOR

The manual review of 150 patients that did not exist in the QOR and were also classified as "Not Cancer Notifiable" by Medtex revealed that there was 1 patient whose reports did suggest a cancer notification. Upon consultation with the QCR, it was confirmed that the patient case would indeed be not notifiable. As a result, no adjustments were required for patients within this stratum where the patient did not exist in the QOR and Medtex patient is not notifiable.

3. Analysis of Medtex patient not notifiable classification and patient exists in the QOR

The manual review of 350 patients that exist in the QOR but were classified as "Not Cancer Notifiable" by Medtex revealed that there were:

- 9 patients whose reports did suggest the need for a cancer notification,
- 6 patients whose reports contained supplementary-only material for cancer notifications, and
- 335 patients whose reports did not suggest the need for a cancer notification.

An analysis on the 335 patients in the QOR with non-notifiable reports revealed that 30 patients were diagnosed in 2009 but had non-notifiable 2009 reports due to patients being either 1) from the private sector or from another Australian state, 2) having cancer notifications from other sources beyond histology and cytology reports, for example, hospital notifications, and 3) having incomplete records due to the inaccessibility of reports occurring in early 2009 due to pathology feeds only being fully operational from mid-March 2009. The remaining 305 patients in the QOR with non-notifiable 2009 reports were cancers diagnosed in other years. Due to the 335 true negative cases, there is an adjustment rate of 335/350 = 0.9571 (95% CI: 0.9287-0.9749). Extrapolating these true negative findings within the respective patient stratum revealed that 10,549 patients (i.e., adjustment rate x 11,021; 95% CI: 10,235-10,744) were correctly classified as Medtex not notifiable; thus requiring adjustment.

With regards to the 9 patients where Medtex did misclassify reports as "Not Cancer Notifiable", these errors were observed to broadly fall into 3 categories, namely, 1) Free text to SNOMED CT mapping errors, 2) Incorrect implementation of Cancer Registry business rules, and 3) other algorithm related issues from Medtex such as the incorrect application of negative assertion phrases to SNOMED CT concepts. These algorithmic issues can be the target of future refinements and therefore may be possible to rectify and resolve through the updating of business rules and improvements to the Medtex free text to SNOMED CT mapping engines and negation detection algorithms. Due to the 9 false negative cases, there is an error rate of 9/350 = 0.0257 (95% CI: 0.0126-0.05). Extrapolating these false negative findings within the respective patient stratum revealed that 283 patients (i.e., error rate x 11,021; 95% CI: 139 to 551) were Medtex not notifiable but do contain cancer. As these cases are within the correct patient stratum where Medtex classified a patient incorrectly as "Not Cancer Notifiable", no further adjustments were required for these errors. These cases may be considered as high importance as these represent missed cancer cases. But as mentioned above, these errors can be potentially rectified or resolved through further system improvements.

4. Analysis of Medtex patient notifiable classification and patient not exist in the QOR

The manual review of 350 patients that do not exist in the QOR but were classified as "Cancer Notifiable" by Medtex revealed that there were:

[1] 95% confidence interval calculated using Wilson's procedure (including continuity correction; http://www.vassarstats.net/prop1.html)

- 82 patients whose reports did suggest the need for a cancer notification,
- 23 patients whose reports only contained supplementary-only material for cancer notifications (as well as another 68 classifications where both Medtex and the cancer clinical coder assigned patients as "Supporting Notifiable Report"), and
- 125 patients whose reports did not suggest the need for a cancer notification.

An error analysis on the 125 patients that Medtex misclassified as "Cancer Notifiable" revealed that the majority of cases were misclassified into particular tumour streams such as skin (misclassifying melanoma and SCC/BCC of skin; Medtex was unable to ascertain whether the SCC/BCC was associated with "skin"), brain and blood cancers. Due to the 125 false positive cases, there is an error rate of 125/350 = 0.3571 (95% CI: 0.3073- 0.4101). Extrapolating these false positive findings within the respective patient stratum revealed that 1,135 patients (i.e., error rate x 3,178; 95% CI: 977 to 1303) were Medtex notifiable but do not contain cancer. As these cases are within the correct patient stratum where Medtex classified a patient incorrectly as "Cancer Notifiable", no further adjustments are required for these errors.

Regarding the 82 patients who were not in the QOR but were indeed cancer notifiable patients, 9 of the patients were recorded in the QOR with different names and therefore could not be computationally matched. Of the remaining 73 patients who could not be found in the QOR, 52 cases related to leukemia, prostate and cervix cancers. Others were distributed with low frequency counts across other tumour streams. In terms of histological type, 57 out of the 73 patients had cancers in the following four morphologies: SCC, CIN 3, adenocarcinoma and chronic lymphocytic leukaemia, B-cell type. Others were distributed with low frequency counts across other morphologies. Due to the 82 true positive cases, there is an adjustment rate of 82/350 = 0.2343 (95% CI: 0.1916-0.2829). Extrapolating these true positive findings within the respective patient stratum revealed that 745 patients (i.e., adjustment rate x 3,178; 95% CI: 609 to 899) were Medtex notifiable and patients were in fact "Cancer Notifiable". Note that 663 (i.e., 73/350 x 3,178; 95% CI: 534 to 813) out of the 745 (95% CI: 609 to 899) "Cancer Notifiable" patients were not found in the QOR. These cases may be considered as high importance as these cases result in unreported cancers.

Adjusted contingency table and overall system performance

In light of the analysis of the comparisons between the QOR, Medtex, and an expert clinical coder, adjustments to Table 1 were required in order to reflect the true performance of Medtex. In particular, the following adjustments were required:

- 682 (95% CI: 320 to 1356) "Cancer Notifiable" patients identified by Medtex but were "Not Cancer Notifiable"
- 10,549 (95% CI: 10,235-10,744) "Not Cancer Notifiable" patients identified by Medtex who are in the QOR but did not have cancer notifiable 2009 pathology reports, and
- 745 (95% CI: 609 to 899) "Cancer Notifiable" patients identified by Medtex who were not in the QOR but indeed had cancers.

Furthermore, to take into account of "Supporting Notifiable Report" classifications, which may over estimate the number of "Cancer Notifiable" cases during the patient-level consolidation process by Medtex, this effect was accounted for by further adjusting the contingency table, especially in the patient stratum where Medtex classified a patient as "Cancer Notifiable" but the patient did not exist in the QOR; here there were 68 out of the 350 cases where both Medtex and the cancer clinical coder assigned patients as "Supporting Notifiable Report" within the respective patient stratum. As a result, there is an adjustment rate of 68/350 = 0.1943 (95% CI: 0.155-0.2405). Extrapolating this result within the respective patient stratum reveals that 617 patients (i.e., adjustment rate x 3,178; 95% CI: 493 to 764) were Medtex "Cancer Notifiable" but were technically "Supporting Notifiable Reports". A further adjustment is therefore applied to the contingency table:

- 617 (95% CI: 493 to 764) Medtex "Cancer Notifiable" patients who were not in the QOR but are technically "Not Cancer Notifiable" due to "Supporting Notifiable Report" classifications, indicating either suspected or re-excisions with no residual cancers.

Table 3 presents the adjusted contingency table and Table 4 presents a summary of Medtex' performance in terms of sensitivity, specificity, positive predicted value and F1 score.

Table 3. Adjusted contingency table of results comparing Medtex and the QOR based on analysis of discrepancies from a manual review*.

		Medtex		Total
		Patient Notifiable	Patient Not Notifiable	
Virtual QOR	Patient Notifiable	12799 (-682, +745) = **12,862 (TP)** *[Min: 12,052; Max: 13,378]*	11021 (-10549) = **472 (FN)** *[Min: 277; Max: 786]*	13,334
	Patient Not Notifiable	3178 (+682, -745, -617) = **2,498 (FP)** *[Min: 1,835; Max: 3,432]*	58504 (+10549 +617) = **69,670 (TN)** *[Min: 69,232; Max: 70,012]*	72,168
Total		15,360	70,142	85,502

*Numbers reported for QOR are "virtual" numbers extrapolated from an analysis of discrepancies between Medtex and the QOR on 2009 pathology data. *Min* and *Max* are computed from the lowest and highest combinations of 95% confidence interval (CI) uncertainties, respectively.

Table 4. Medtex' performance on classifying cancer notifiable patients*.

Sensitivity	Specificity	Positive Predictive Value	F1 score
96.5% (94.5-97.8%)	96.5% (95.3-97.4%)	83.7% (79.6-86.8%)	89.6% (86.4-91.9%)

*95% confidence interval (CI) uncertainties reported in brackets.

Discussion

Medtex achieves high sensitivity and specificity in identifying cancer notifiable patients from histology and cytology HL7 reports in line with the results from previous work using a smaller and unrepresentative dataset[9]. As a consequence of Medtex' high sensitivity, the system was able to identify cancer patients who were not found in the QOR. The high sensitivity, however, was at the expense of PPV. Although the effect of "Supporting Notifiable Report" classifications on false positives was partially taken into account, the lower PPV is considered to be of lower priority (or importance) as these cases can be quickly reviewed manually. The observed lower PPV finding was only possible through this larger scale study with a larger range of report types, tumour streams, pathology laboratories and pathologists; previous work on a smaller dataset did not reveal this[9]. Error analysis reveals that system errors (both false negative and false positives) were mainly tumour stream dependent and therefore could be targeted as part of future work. Future work would also include the evaluation of cancer notifications at a cancer case level to ensure that "Cancer Notifiable" patients were correctly matched to their specific cancers.

Unlike other approaches to cancer notifiable classification that rely on custom phrases, explicitly mentioned disease codes, or the development of tumour specific classification models, Medtex aims to be generic and relies on Cancer Registry business rules and the normalisation and reasoning of records using standard clinical terminologies, specifically SNOMED CT. Although, the symbolic rule-based approach used in Medtex was developed from a small development data set, the current study shows how the utility of such a system can be assessed when applied on larger datasets. The findings from this larger dataset study can be used to bootstrap the development of the symbolic rule-based system. Machine-learning approaches to classification require larger training datasets to generalise but labelled data can be difficult to obtain especially from busy and highly skilled health personnel.

The Medtex processing of reports can also be applied at pathology laboratories to assist with fulfilling mandatory cancer notification requirements. Medtex could be used by pathology laboratories to accurately and automatically flag cancer reports for notification to cancer registries. Not only would this benefit cancer registries, but also individual pathology laboratories; thus improving cancer notification management.

Medtex also has the added advantage of producing uniform data, reducing manual processing and assisting in preventing inaccuracies in the data. This can allow cancer registries and pathology laboratories to devote staffing resources to other high value adding areas such as quality assurance, data integrity and manage the registry in a cost effective manner.

Conclusion

The utility of a medical text analytic system, Medtex, on automating Cancer Registry notifications from narrative histology and cytology HL7 reports from an Australian state-wide pathology repository was assessed. The system was developed using Cancer Registry business rules, natural language processing and symbolic reasoning using the SNOMED CT ontology. Cancer report notifications were consolidated at a patient level and compared against patients with notifiable cancers from the Queensland Oncology Repository (QOR).

A stratified random sample of 1,000 patients was manually reviewed by an expert clinical coder to analyse agreements and discrepancies. Upon extrapolation of the results to the full dataset, sensitivity of 96.5% (95% CI: 94.5-97.8%), specificity of 96.5% (95.3-97.4%), positive predictive value of 83.7% (79.6-86.8%), and an F1 score of 89.6% (86.4-91.9%) were achieved for identifying cancer notifiable patients. Medtex was also able to identify cancer patients that were not found in the QOR; resulting in unreported cancers. Error analysis revealed that system errors tended to be tumour stream dependent with specific cancer sites and histological types causing a large portion of the errors. The evaluation of Medtex on large-scale medical records enabled insights for assessing the utility of the medical text analytic system as well as highlighting avenues for future system developments.

The system is currently being piloted within the Department of Health, Queensland Government, to process the historical and new incoming electronic HL7 pathology reports, not only for identifying cancer notifiable cases, but also the automatic extraction of cancer notifications data including cancer stage.

The Medtex processing of other sources of cancer notification including feeds from private pathology laboratories, radiology reports and death certificates is also currently being investigated. Medtex can also be applied to other medical domains beyond cancer.

Acknowledgements

This research was done in partnership between the Australian e-Health Research Centre (AEHRC) within CSIRO and the Queensland Cancer Control Analysis Team (QCCAT) within the Department of Health, Queensland Government. The authors acknowledge Stephen Armstrong and Nilesh Mendis from QCCAT for data access as well as help in getting Medtex deployed on the Department of Health's infrastructure.

References

1. Queensland Government. Cancer in Queensland: A Statistical Overview 1982-2021, Annual update 2012. Queensland Health, Brisbane, 2015
2. Hanauer DA, Miela G, Chinnaiyan AM, et al. The registry case finding engine: an automated approach for cancer patient identification from unstructured, free-text pathology reports. *J Am Coll Surg* 2007;205(5):690-7.
3. Contiero P, Tittarelli A, Maghini A, et al. Comparison with manual registration reveals satisfactory completeness and efficiency of a computerized cancer registration system. *J Biomed Inform* 2008;41(1):24-32.
4. D'Avolio LW, Nguyen TM, Farwell WR, et al. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). *J Am Med Inform Assoc* 2010;17(4):375-82.
5. Osborne JD, Wyatt M, Westfall AO, et al. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. *J Am Med Inform Assoc* 2016. DOI:10.1093/jamia/ocw006
6. Dale D, Golabek JK, Chong N. The impact of E-path technology on Ontario Cancer Registry operations. J Registry Manag. 2002;29:52-56.
7. Brueckner P. Automated identification and coding of cancer pathology reports. Advancing Practice, Instruction, and Innovation through Informatics (APIII 2005): Scientific Poster Session Abstracts. Archives of Pathology & Laboratory Medicine: June 2006;130(6):890-903.
8. Nguyen AN, Lawley MJ, Hansen DP, et al. A Simple Pipeline Application for Identifying and Negating SNOMED Clinical Terminology in Free Text. Health Information Conference 2009:188-196.
9. Nguyen A, Moore J, Zuccon G, et al. Classification of pathology reports for Cancer Registry notifications. *Stud Health Technol Inform* 2012;178:150-6.
10. Nguyen A, Moore J, Lawley M, et al. Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications. *Stud Health Technol Inform* 2011;168:117-24.
11. Nguyen A, Moore , O'Dwyer J, Philpot S. Assessing the Utility of Automatic Cancer Registry Notifications Data Extraction from Free-Text Pathology Reports. AMIA 2015 Annual Symposium, 2015.
12. Nguyen AN, Lawley MJ, Hansen DP, et al. Symbolic Rule-based Classification of Lung Cancer Stages from Free-Text Pathology Reports. *J Am Med Inform Assoc* 2010;17(4):440-5.

13. Nguyen A, Lawley M, Hansen D, et al. Structured pathology reporting for cancer from free text: lung cancer case study. *electronic Journal of Health Informatics* 2012;*7*(1):e8.
14. Cunningham H, Maynard D, Bontcheva K, et al. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Annual Meeting of the ACL, 2002.
15. Richards M, Monson-Haefel R, Chappell DA. Java Message Service (2nd ed.). O'Reilly Media, Inc 2009
16. Queensland Cancer Registry. Clinical Coding Manual Version 3.