Title:

Collection of Cancer Stage Data by Classifying Free-text Medical Reports

Authors and Affiliations:

[1]Iain A. McCowan (PhD), [1]Darren C. Moore (MEng), [1]Anthony Nguyen (PhD), [2]Rayleen V. Bowman (PhD, FRACP), [3]Belinda E. Clarke (PhD, FRCPA), [3]Edwina E Duhig (FRCPA), [4]Mary-Jane Fry.

[1]CSIRO e-Health Research Centre, Brisbane, Australia.

[2]Department of Medicine, University of Queensland, Brisbane, Australia.

[3]Department of Anatomical Pathology, The Prince Charles Hospital, Brisbane, Australia.

[4]Queensland Cancer Control Analysis Team, Queensland Health, Brisbane, Australia.

Corresponding Author (for all communications, including reprints requests):

Iain McCowan

e-Health Research Centre

PO Box 1084 Adelaide Street

Brisbane QLD 4000

AUSTRALIA

Tel: +61 7 3024 1626

Fax: +61 7 3024 1690

Email: iain.mccowan@csiro.au

Abstract

Cancer staging provides a basis for planning clinical management, but also allows for meaningful analysis of cancer outcomes and evaluation of cancer care services. Despite this, stage data in cancer registries is often incomplete, inaccurate or simply not collected. This article describes a prototype software system (Cancer Stage Interpretation System, CSIS) which automatically extracts cancer staging information from medical reports. The system uses text classification techniques to train support vector machines (SVM) to extract elements of stage listed in cancer staging guidelines. When processing new reports, CSIS identifies sentences relevant to the staging decision, and subsequently assigns the most likely stage. The system was developed using a database of staging data and pathology reports for 710 lung cancer patients, then validated in an independent set of 179 patients against pathologic stage assigned by two independent pathologists. CSIS achieved overall accuracy of 74% for tumour (T) staging and 87% for node (N) staging, and errors were observed to mirror disagreements between human experts.

Keywords

Cancer Staging; Lung Cancer; Machine Learning; Clinical Decision Support Systems.

I. INTRODUCTION

Cancer stage categorises the size and location of the primary tumour, the extent of lymph node involvement, and presence or absence of metastatic spread to other body parts. The clinical management of most cancers according to evidence-based guidelines (e.g. [1]) is dependent upon the stage of disease at diagnosis, and documentation of cancer stage at diagnosis is increasingly being recommended as a standard of care by national cancer bodies. International standards for cancer staging have been developed, such as the TNM (Tumour Nodes Metastases) standard defined by the AJCC (American Joint Committee on Cancer) and UICC (International Union Against Cancer), summarised in Table 1 [2].

T: Primary Tumour	X	Primary tumour cannot be assessed.
	0	No evidence of primary tumour.
	is	Carcinoma in situ.
	1,2,3,4	Increasing size and/or local extent of the primary tumour.
N: Regional Lymph Nodes	X	Regional lymph nodes cannot be assessed.
	0	No regional lymph node metastasis.
	1,2,3	Increasing involvement of regional lymph nodes.
M: Distant Metastasis	X	Distant metastasis cannot be assessed.
	0	No distant metastasis.
	1	Distant metastasis.

Table 1: Summary of the TNM staging protocol [2].

Apart from the important role played by cancer staging in the clinical management of individual patients, there is increasing acknowledgement that outcomes analysis of cancer management or intervention programs on a population, governance or facility level is meaningful only if interpreted in the light of this major prognostic factor. As the main population based data repositories, cancer registries have moved to incorporate clinically relevant fields such as cancer stage, in order to enable more accurate and useful outcomes analysis. Despite these changes, stage data in registries is still commonly absent or incomplete. After four years of mandated stage data collection for prostate cancer by the

Maryland Cancer Registry, data was still missing in 13% of cases on average, and up to 20% in some regions [3]. A similar study in the Ottawa Regional Cancer Centre found missing staging information in 10% of lymphoma cases, and 38% of breast cancer cases [4]. An earlier study at that centre showed that mandated stage data collection across all cancer types led to complete stage data being available for 71% of cases on average [5]. Organised stage data collection as undertaken in these two North American centres is in contrast to many other regions. For instance, in 2004 the National Cancer Control Initiative reported that there was no ongoing population-based collection of staging information in any Australian state or territory [9].

Even when collected, there is evidence that stage data is often inaccurate. A study of demographic differences in prostate cancer staging in Connecticut found that 23% of cases in the registry were incorrectly coded [6], either due to incomplete medical records or staging errors. A review of lung cancer stage data in the Maastricht Cancer Registry in the Netherlands found major discrepancies in 12% and minor discrepancies in 23% of cases [7]. Many of these were due to incorrect application of staging guidelines, as well as data entry errors. Similarly, a review of stage data in Ottawa Regional Cancer Centre found staging errors occurred in 2-5% and data entry error in 3-6% of all cases [5]. There were differences between registry stage and stage as determined from available clinical information in 31% of lymphoma and 8% of breast cancer cases [4].

When not obtained directly from clinicians prospectively, it is possible to perform retrospective staging based on retrieved medical records. A Nottingham prostate cancer study which retrospectively assigned stage using case notes showed that stage information regarding the primary tumour (T stage) could be abstracted for 96% of cases, however only limited information was available for staging lymph node and metastatic involvement (N and M

stage) [8]. The Western Australian Cancer Registry feasibility study of staging from medical records for 20 cancer types found that, under various assumptions, stage data could be collected using current data sources for 7 cancer types, but was not feasible or required system change for the others [9]. The same group subsequently undertook a project to retrospectively collect stage data for all colorectal cancer cases over a one year period [10]. They were able to fully stage 76% of cases from available data sources (pathology reports, case notes, hospital registries, etc), and a further 22% of cases if M stage was omitted. A study in which stage data was retrospectively sourced from medical reports was used to monitor cancer outcomes for indigenous Australians in the Northern Territory [11]. Therefore, while staging is a recognised component of providing quality cancer care, data on stage is often incomplete, inaccurate or not recorded. Furthermore, while it is possible to retrospectively retrieve data from available medical reports, doing this manually can be time and labour intensive.

Motivated by these limitations, we developed CSIS (Cancer Stage Interpretation System), a prototype software system to assign cancer stage data by automatically extracting relevant information from free-text medical reports stored in clinical information systems. CSIS could be used by a cancer registry to support collection of staging information for those patients not formally staged by human experts, allowing more comprehensive population-level analysis of outcomes. Alternatively, if deployed at the point of reporting, it has potential to improve the efficiency and consistency of staging by clinicians. While the system was developed on lung cancer data available to us, it could in principle be applied to stage other cancer types. For an individual patient, input to the system consists of textual reports describing pathology tests. The objective is to estimate pathologic stage by applying machine learning text categorisation techniques [12]. As metastatic lung cancer is defined as involvement of other organs, it is not

usually assessable from pathological studies of the lung, therefore the current system does not attempt to determine the M stage.

Previous work investigated direct classification of the cancer stage using binary SVMs operating on the concatenated reports of a given patient [13], [14], essentially posing the problem as document-level topic categorisation [12]. While results were promising, there was a need to improve system performance. Furthermore, the direct report-level stage classification meant it was not possible to detail reasons for the stage classification, which was desirable to interpret errors and build user trust. Traditional topic categorisation models a document as a collection of words representing a number of topics. While this is an appropriate model for tasks such as news report categorisation, it does not well-fit the current task. A better model of medical reports is a sequence of specific statements relating to different diagnostic factors. With this motivation, the system proposed in this article instead determines the stage indirectly, by first determining the presence or absence of specific staging factors using sentence-level classifiers. The staging protocol, such as shown in Table 1, is then applied to assign the most advanced stage associated with a positive finding. As well as potentially improving the accuracy of the eventual stage assignment, decomposing the stage in this way declares reasons behind the decision, linked to the supporting sentences.

The remainder of this article is organised as follows. Section II presents a review of related work, covering text categorisation and software support for cancer staging. Section III then gives an overview of the proposed automatic method for collecting cancer stage data and implementation details. Results of a system trial evaluation are presented in Section IV, followed by discussion of the major findings in Section V. Finally, concluding remarks and directions for future research are presented in Section VI.

II. BACKGROUND

This paper presents a system to automatically extract cancer stage information using text categorisation techniques. Other researchers have previously presented automatic cancer staging algorithms using high-level structured input data coding major diagnostic factors for cervical, ovarian and prostate cancer [38], [39], [40], [41], [42], [43]. Other than these automatic methods, software has been developed for staging with synoptic data entry forms [44], [45], as well as converting between different staging protocols [45].

In previous work, we reported a novel approach to automatic staging by direct report-level classification of the stage from free-text histology reports using SVM's and a bag-of-words representation [13],[14]. By using available free-text reports rather than relying on expert coding, the approach allowed for broader applicability than previous staging software, particularly for retrospective data collection and when access to expert knowledge of staging is limited. A review of the literature on medical text categorisation was presented in [13], and is summarised here. Traditionally, text categorisation is the task of determining if a given document belongs to each of a predefined set of classes [12]. Most recent research has concentrated on machine learning approaches which automatically build classifiers by learning the characteristics of each category from a set of pre-classified documents [12], [15]. These most commonly use a bag-of-words document representation and Support Vector Machine classifiers (SVM's) [16] [17], although many other classifiers have been investigated [18], [19], [20], [12]. Within the medical domain, a number of comparative studies have demonstrated that SVM's outperform other classifier types across a range of medical text classification tasks (e.g. [24], [28], [34]).

The system proposed in the present article builds on prior work [13],[14] by determining the presence or absence of specific staging factors using a two-level sentence classification

approach. For each factor from the staging guidelines, sentences are first classified for relevance and then as either a positive or negative finding. There has been little prior work in the text classification literature where the unit of classification is smaller than the entire document, however related approaches have been proposed for extractive text summarisation. Such systems generally include a step where a subset of important sentences are classified from a document using various features, such as sentence length or location, as well as term frequencies [46], [47]. A double classification methodology, in which sentences are first classified as containing relevant information or not, and then terms of interest are classified from within these relevant sentences, was proposed in [49]. In other related work, sentences from MedLINE abstracts were accurately categorised according to four types using a sentence-level bag-of-words SVM in [48].

III. SYSTEM DESCRIPTION

A Architecture

Figure 1 shows the high level architecture of the proposed system. The system components are described in the following subsections. CSIS is a prototype software system implementing these components in a command-line application, which inputs a list of patients with corresponding free-text report files, and outputs an XML file with the derived staging metadata. More specific implementation details, such as SVM training methods, follow in Section III.B. The system employs a text pre-processing stage to standardise report texts, followed by support vector machine (SVM) T and N relevance classifiers that assess the relevance of each report to staging tasks. Sentences from relevant reports are then each analysed by a series of SVM- and rule-based classifiers corresponding to specific contributing factors defined in the staging guidelines. Sentence level classifier results are post-processed to determine the final T and N stage.

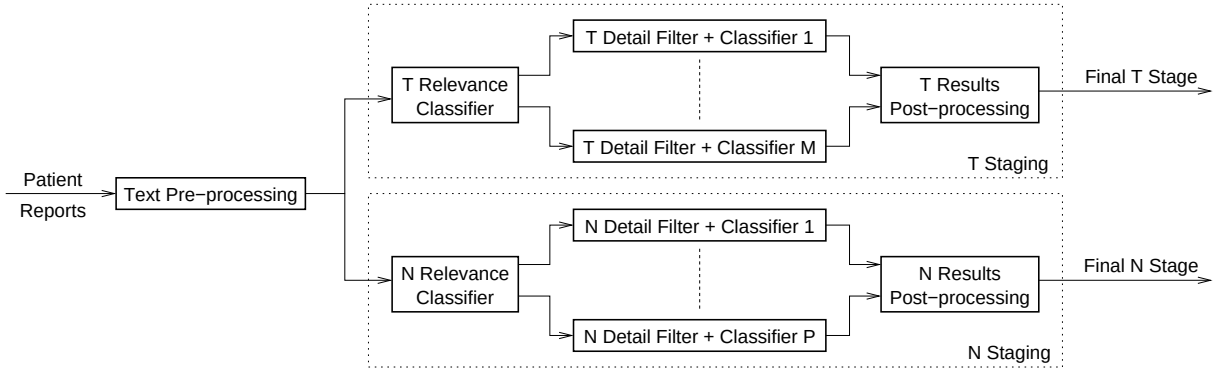

Figure 1: Proposed system high level architecture

Step 1	There	is	no evidence of	lymph node	metastases
Step 2	There	is	_PRENEG_	lymph node	metastases
Step 3	E0060550	E0012152	_PRENEG_	E0038425	E0039864
Step 4	E0060550	E0012152		NEG_E0038425	NEG_E0039864

Table 2: Example output of each text pre-processing step. In this case, the original 8-word sentence is mapped into a sequence of 4 input terms for subsequent classification.

1) Text Pre-processing

The purpose of text pre-processing is to standardise the report texts and to decrease variability by encoding common terms or phrases using a biomedical dictionary, the Unified Medical Language System (UMLS) SPECIALIST Lexicon [50]. The text pre-processing system in the current system is based on that reported in [13], and consists of four steps: 1) Normalisation, 2) Detection of negation phrases, 3) Conversion to UMLS SPECIALIST term codes, and 4) Negating relevant terms. The steps implementing normalisation and conversion to UMLS SPECIALIST term codes are described in [13]. As in the prior work, the NegEx algorithm [31], [32] was used to detect negation phrases. In the current system, the list of approximately 1400 terms considered for negation in Step 4 comprised terms occurring in at least 5 reports in the development data set. Negation phrase codes inserted in Step 2 are removed after they have been applied to surrounding terms. Table 2 shows example output of each step.

2) Report Relevance Classification

Pathology reports for lung cancer often contain insufficient macroscopic detail to enable T or N staging. Most reports on small lung biopsy samples, where the

emphasis is on microscopic findings, fall into this category. The purpose of relevance classification is to identify which of a patient's reports are not useful for T or N staging so they are excluded in subsequent steps. If all reports for a patient are classified as irrelevant to T or N staging, then the patient is automatically assigned a stage of TX or NX, respectively. T and N report relevance classifiers are implemented using SVMs that classify a bag-of-words representation of each report.

3) Stage Detail Classifiers

The proposed classification strategy (see Figure 1) uses sentence level classifiers corresponding to specific factors from the staging guidelines. Examples of factors influencing a T stage assignment are the maximum dimension of the tumour and whether it invades the main bronchus or chest wall. Factors that affect a particular N stage assignment are related to tumour involvement of particular anatomical lymph node groups (e.g. peribronchial, mediastinal, etc).

The system starts with a default stage of T1/N0 (and thus assumes patients are known to have lung cancer) and upgrades this to the highest stage associated with any of the factors classified as positive across all sentences for that patient. Factors classified as negative are not explicitly taken into account when assigning the final stage.

Table 3 lists the sentence level classifiers that were implemented along with their type. All classifiers employ keyword filtering as a first step to eliminate entirely unrelated sentences (e.g. a sentence must contain a dimension in order for it to be input to the tumour size classifier). Most sentence-level classifiers use a 2-level SVM approach, in which a first level SVM classifies a sentence as being either relevant (i.e. supports either a positive or negative finding) or irrelevant to the factor

in question. A relevant sentence is then classified as supporting a positive or negative finding by the second level SVM.

Individual 2-level SVM classifiers were implemented for tumour size (TS) as well as for each type of lymph node involvement (PLN, HLN, MLN, SCLN – Table 3 defines these classifier short names). For staging factors related to invasion of body sites by the primary tumour, a common "invasion" classifier was implemented. Each sentence is pre-processed to convert the UMLS SPECIALIST lexicon term representation of relevant body parts (e.g. visceral pleura, chest wall, etc) to a common "_BODYPART_" term. A similar transformation is applied to tumour terms (e.g. mass, lesion → _TUMOUR_) and to terms/phrases implying invasion (e.g. involves, extends into → _INVADE_). Each transformed sentence is then input to a common 2-level SVM classifier as described above. For sentences classified as positive by the 2-level SVM, a rule-based post-processing step examines the untransformed version of the sentence to discover the particular body part that is undergoing invasion by the primary tumour.

Due to lack of positive examples in development data set, the SEPN (defined in Table 3) was implemented as a rule-based classifier that searches for phrases implying the existence of secondary tumour deposits in the same lobe. Similarly, the NONM searches for blanket statements commonly used by reporting pathologists to indicate that no lymph nodes are involved by the cancer. A positive finding from this classifier overrides the other N stage factor decisions.

Classifier	**Short Name**	**Stage Association**	**Classifier Type**

Max. tumour dimension ≤ or > 3cm	TS	T2	2-level SVM
Visceral pleural invasion	VP	T2	Invasion SVM*
Main bronchus invasion	MB	T2	Invasion SVM*
Chest wall invasion	CW	T3	Invasion SVM*
Diaphragm invasion	DIA	T3	Invasion SVM*
Mediastinal pleural invasion	MEDP	T3	Invasion SVM*
Parietal pericardium invasion	PPER	T3	Invasion SVM*
Great vessel invasion	GV	T4	Invasion SVM*
Mediastinum/heart/trachea/oesophagus/ visceral pericardium invasion	T41	T4	Invasion SVM*
Vertebral body/carina/vagus nerve invasion	T42	T4	Invasion SVM*
Separate tumour nodules in same lobe	SEPN	T4	Key-phrase
No nodal involvement	NONM	N0	Key-phrase
Peribronchial lymph node involvement	PLN	N1	2-level SVM
Hilar lymph node involvement	HLN	N1	2-level SVM
Mediastinal lymph node involvement	MLN	N2	2-level SVM
Subcarinal lymph node involvement	SCLN	N2	2-level SVM

Table 3: List of sentence level classifiers used in the proposed system (* indicates common 2-level SVM classifier with post-processing to determine factor)

B System Development

1) Development Corpus

To train and validate the system, a corpus of de-identified medical reports with corresponding pathological staging data was obtained following research ethics approval. The pathological staging data was obtained from a database [1] collected over the five year period ending in December 2005. The corresponding medical reports were extracted from a pathology information system. A total of 8 cases from the available data sources had pathological stages of T0, TX, Tis, or N3. Automatic classifiers for these stages were therefore not implemented and cases with those

stages were omitted from the corpus. The development corpus statistics are included in Table 4.

Data	Cases	Stage Breakdown				Reports
Pathology reports + pTNM	710	T1	204	NX	57	817
		T2	405	N0	432	
		T3	52	N1	149	
		T4	49	N2	72	

Table 4: Key statistics for the development data set

Training sets for the report relevance classifiers described in Section III.A2) were derived from the development corpus by annotating each of the reports with a relevant/irrelevant label for both T and N staging.

A separate training set was derived from the development corpus for each of the sentence level factor classifiers described in Section III.A3) by splitting all reports into individual sentences, filtering out irrelevant sentences using the keyword filter for that classifier, and then annotating remaining sentences with one of three labels (irrelevant, -ve finding, +ve finding). Note that direct classification of NX was not done as this result was derived from the N report relevance classifier output.

2) SVM Implementation

The bag-of-words term weights used for text representation with all SVMs throughout the baseline and proposed systems were calculated according to the LTC-weighting scheme [51]. The LTC weighting is commonly used in state-of-the-art text categorisation systems, as it effectively de-emphasises common terms (occurring often in many documents), produces normalised weights across different length documents, and reduces the impact of large differences in frequency through use of the logarithm.

A common training strategy was used with all SVM-based classifiers. A cross-validation approach was used to optimise SVM training parameters and decision threshold and to obtain unbiased classifier output over the entire development training set. The SVM^{light} [52] toolkit was used for all SVM training and testing. The optimal parameters discovered through cross-validation were used to train a final classifier on all training data. Decision thresholds were selected by cross validation to equalise sensitivity and specificity. No attempt was made to adjust individual classifier decision thresholds to optimise the global T and N staging accuracy.

3) Development Results

Unbiased classifier outputs from the report relevance classifiers and the sentence level staging factor classifiers described above were merged to obtain final T and N staging results on the 710-case development set. For T staging, 77.6% correct (95% CI = 74.3-80.6)[1] was obtained for classifying the 5 T stages (TX, T1-T4). For N staging, 81.8% correct (95% CI=78.8-84.6) was obtained for classifying 4 N stages (NX, N0-N2).

To compare the sentence-level classification with the previous direct report-level approach, a multi-class SVM system was used as a baseline. This approach directly classifies T and N stage from a concatenation of reports for each patient, with the multi-class classification implemented as a hierarchy of binary SVMs, and is fully described in [14]. As TX and NX classes were not considered in [14] , to allow direct comparison of results, the baseline system was augmented with the T and N report relevance classification stage from the current proposed system. On the same 710-case development set, baseline system performance was 62.8% (95% CI=59.1-66.4) and 77.0% (95% CI=73.7-80.1) correct for T and N staging respectively. This was used as the baseline system in the trial evaluation reported in Section IV.

[1] All 95% confidence intervals reported in this article are calculated using the Wilson procedure.

These development results indicate that accuracy has been improved by decomposition into sentence-level staging factor classifiers, as opposed to the more conventional document-level approach that directly classifies final T and N stages.

IV. STATUS REPORT

To evaluate the reliability of the proposed system, a trial was conducted, as described in the following sections. Some findings from this trial were presented in preliminary form in [54].

A Trial Objectives

1. *To study the level of agreement in expert staging decisions.* Subjectivity in the stage decision may arise from inconclusive examinations, varying interpretations of staging criteria, or ambiguity in the way the results are communicated. The first objective of the trial was to quantify the degree of variability between two independent human experts.
2. *To evaluate the performance of automatic staging decisions.* The second purpose of the trial was to evaluate the performance of the automatic cancer stage assignment, in comparison to a 'gold standard' based on the same input information. For this purpose the 'gold standard' consisted of stage independently assigned in perfect agreement between two human experts. For the few cases where human experts disagreed, one expert's decision was selected at random as the gold standard.
3. *To evaluate the reliability of classifying key stage factors.* Finally, in addition to overall T and N stage assignments, we evaluated how well the system classified specific factors based on key sentences in relation to the human experts.

B Method

1) Input Data

The trial data set consisted of pathology reports for lung cancer cases that were not seen during the development phase, and was extracted from the same pathology information system as the development data set. The trial set consisted of reports for

116 cases that had been assigned a formal pathologic stage in the eight month period subsequent to December 2005, along with 63 unstaged cases that had a report describing examination of a lung or lobe from a pneumonectomy or lobectomy procedure.

2) Output Data

Two expert pathologists competent in lung cancer staging were presented with the de-identified reports for the 179 patients. They then independently classified the TNM stage and specific factors (from Table 3) for each patient, and entered the data into an electronic form. Form validation required the pathologists to enter T and N stages, however default values were set for all other data fields (M stage of "MX", and "negative" for all other details). A text box was also provided on the form to allow any free-text comments to be entered.

In order to determine the 'gold standard' TNM stage for system evaluation, following independent data collection from the pathologists, a meeting was convened to discuss cases where the experts differed in the assigned TNM stage. In this meeting, a consensus TNM stage was assigned by the experts for as many cases as possible. If consensus was not reached for a case, due to ambiguity in the report language or staging guidelines, the two different TNM stages were retained.

System output consisted of the T and N stage, along with the output of the detail classifiers from Table 3. To preclude bias, processing of the trial data was performed by technicians independent of the development team investigators, so that investigators were blind to the trial data set.

3) Performance Measures

The following defines the measures used to evaluate results based on the total Number of patients (N), along with counts of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) resulting from classification decisions. To

evaluate the overall performance of the system for T and N staging, multi-class classification performance is measured using the *accuracy*,

$$Acc = \frac{TP}{N}$$

Agreement between human experts was measured by the *kappa* statistic, which takes account of agreement occurring by chance.

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A)=Accuracy is the observed agreement, and

$$P(E) = \sum_{c=1}^{C} \frac{N_1(c)}{N} \frac{N_2(c)}{N}$$

is the agreement expected by chance, where $N_1(c)$ is the number of times annotator 1 selected class *c*. Binary classification performance is measured using the *sensitivity* and *specificity*.

$$Sens = \frac{TP}{(TP + FN)} \; ; \; Spec = \frac{TN}{(TN + FP)}$$

The *confusion matrix* is a 2-dimensional tabulation of frequency counts according to assigned (test) class labels and actual (gold standard) class labels. By highlighting commonly occurring class confusions, the confusion matrix is a useful tool for analysing multi-class classification systems.

C Results

1) Expert Agreement

The inter-expert agreement is shown in Table 5 in terms of the kappa statistic and raw percentage agreement for T and N staging on the complete 179-case trial data set. The break-down of cases by stage is demonstrated in the confusion matrix in Table 6.

Stage	Kappa	% Agreement (95% CI)

T	0.83	89.9 (84.3-93.8)
N	0.96	97.8 (94.0-99.3)

Table 5: Inter-expert agreement for T and N staging.

		Expert 2			
		T1	T2	T3	T4
Expert 1	T1	**49**	0	2	0
	T2	1	**94**	3	2
	T3	0	3	**7**	2
	T4	0	5	0	**11**

		Expert 2			
		NX	N0	N1	N2
Expert 1	NX	**16**	1	0	0
	N0	0	**107**	1	2
	N1	0	0	**35**	0
	N2	0	0	0	**17**

Table 6: Confusion matrices comparing T and N stage assigned by Experts 1 and 2.

The confusion matrices show there were 18 T stage and 4 N stage disagreements between the experts. In the subsequent meeting to determine gold standard stage data for system evaluation, as described in Section IV.B2), the experts were able to reach consensus for 10 of the 18 T stage cases and all 4 of the N stage cases. Two different T stage assignments were retained for the remaining 8 cases.

The inter-expert agreement for each of the detailed staging factors is shown in Table 7. A Kappa value of N/A (Not Applicable) indicates no instances found by experts (division by zero). Numbers for advanced T stage factors were small, so the significance of results for these factors is not clear.

		Expert 1 vs Expert 2				
		Agree		**Disagree**		
Stage	**Classifier**	**YY**	**NN**	**YN**	**NY**	**Kappa**
T	TS	96	80	1	2	0.97
	VP	63	109	2	5	0.92
	MB	0	173	6	0	0.00
	CW	5	171	2	1	0.76

	DIA	0	179	0	0	N/A
	MEDP	0	176	3	0	0.00
	PPER	0	178	0	1	0.00
	GV	0	179	0	0	N/A
	T41	2	170	1	6	0.35
	T42	0	179	0	0	N/A
	SEPN	8	161	5	5	0.59
N	NONM	76	71	19	13	0.64
	PLN	27	146	6	0	0.88
	HLN	23	150	2	4	0.87
	MLN	12	165	2	0	0.92
	SCLN	6	172	1	0	0.92

Table 7: Inter-expert agreement for detailed staging factors (see Table 3 for classifier name definitions).

2) System Performance

Performance was evaluated against the 'gold standard' of human expert stage assignments. As described above, the experts reached consensus on the T stage in only 171 cases. N stage consensus was attained for all 179 cases. For each of the remaining 8 cases without T stage consensus, one of the expert stage decisions was selected at random as the gold standard. Overall T and N stage accuracy with respect to the expert staging for a baseline system that classifies the stage directly from the concatenated reports for each patient, as described in Section III.B3), as well as for the proposed CSIS is shown in Table 8. The break-down of cases by stage is demonstrated in the confusion matrix in Table 9.

Stage	Cases	Accuracy % (95% CI)	
		Baseline	Proposed
T	179	62.6 (55.0-69.6)	74.3 (67.1-80.4)
N	179	76.5 (69.5-82.4)	86.6 (80.5-91.1)

Table 8: Accuracy of system with respect to experts for T and N stage.

		System			
		T1	T2	T3	T4
Experts	T1	**39**	10	0	1
	T2	6	**80**	2	13
	T3	0	7	**2**	1
	T4	1	5	0	**12**

		System			
		NX	N0	N1	N2
Experts	NX	**10**	6	1	0
	N0	2	**105**	0	1
	N1	0	8	**27**	1
	N2	0	3	2	**13**

Table 9: Confusion matrix comparing T and N stage assigned by Experts and the proposed System[2].

The performance of CSIS on cases with perfect expert agreement[3] for each of the detailed staging factors in terms of sensitivity, specificity, accuracy and kappa statistic is shown in Table 10. Again, the significance of results for advanced T factors is not clear due to low numbers of positive examples.

		Experts vs System							
		Agree		**Disagree**					
Stage	**Classifier**	**YY**	**NN**	**YN**	**NY**	**Sens.**	**Spec.**	**Acc.**	**Kappa**
T	TS	93	67	3	13	0.97	0.84	0.91	0.81
	VP	55	96	8	13	0.87	0.88	0.88	0.74
	MB	0	171	0	2	1.00	0.99	0.99	0.00
	CW	3	170	2	1	0.60	0.99	0.98	0.66
	DIA	0	179	0	0	1.00	1.00	1.00	N/A
	MEDP	0	176	0	0	1.00	1.00	1.00	N/A
	PPER	0	178	0	0	1.00	1.00	1.00	N/A
	GV	0	178	0	1	1.00	0.99	0.99	0.00
	T41	0	170	2	0	0.00	1.00	0.99	0.00
	T42	0	179	0	0	1.00	1.00	1.00	N/A
	SEPN	5	148	3	13	0.62	0.92	0.91	0.34
N	NONM	61	67	15	4	0.80	0.94	0.87	0.74
	PLN	23	141	4	5	0.85	0.97	0.95	0.81

[2] As there were no gold-standard cases, TX results are not reported, however that the system was successful in not inserting any false positive TX cases.

[3] It was not feasible to resolve disagreements on detailed staging factors in the post-trial consensus meeting.

	HLN	20	143	3	7	0.87	0.95	0.94	0.77
	MLN	9	162	3	3	0.75	0.98	0.97	0.73
	SCLN	5	171	1	1	0.83	0.99	0.99	0.83

Table 10: Performance of system for classifying detailed staging factors.

A final point regarding system performance is the incurred processing time. For each report, on a single processor 3 GHz Pentium-4 PC, the report-level baseline system required 1.14 seconds, while the sentence-level proposed system required 1.20 seconds. In both cases, the major component was the text pre-processing stage which required approximately 1 second.

V. DISCUSSION

A Trial Objectives.

1. To study the level of agreement in expert staging decisions.

The comparison between the stage assigned by the two experts shows that there is a degree of subjectivity in determining a patient's T and N stage based purely on the available pathology reports, particularly for the T stage decision. Following initial coding, there were 18 disagreements between Experts 1 and 2 for T staging on the full 179 patient set, and 4 disagreements for N staging. The confusion matrices in Table 6 show that the most common confusions were between T2 and T3, and T2 and T4. These findings broadly correspond with agreement levels found in reviews of registry data [4], [5], [6], [7]. Following discussion between the two experts, consensus was reached on 10 of these T stages, and all 4 N stages. The 10 original T stage disagreements were attributed to 6 reports with ambiguous language and 4 interpretation errors. The 4 original N stage disagreements were attributed to 2 data entry errors, 1 interpretation error, and 1 report with ambiguous language. The remaining 8 T stage decisions on which no consensus could be reached consisted of 4 cases where the staging guidelines are ill-defined for distinguishing a single primary tumour from multi-focal tumours (leading to T2M1 and

T4M0 stage confusion, which are however both Group Stage IV), and 4 cases where the report was imprecise regarding tumour extent (leading to T2/T3 confusion).

2. To evaluate the reliability of automatic staging decisions.

CSIS had T stage accuracy of 74% and N stage accuracy of 87% on the trial data. This represents an improvement of approximately 10% over the previous baseline system for both T and N staging. These results are similar to those observed on the development data set (Section III.B3). In general, higher accuracy for N stage as compared with T stage mirrors the trend observed in the expert disagreements, and the CSIS confusions predominantly occurred between the same advanced T stages as for the human experts.

3. To evaluate the reliability of classifying key stage factors.

The results in Table 10 show that agreement between CSIS and the experts for individual key stage factors also follows the same patterns observed between human experts in Table 7. The sentence-level factor classifier results in Table 10 explain the reasons for CSIS stage errors. Confusion between T1 and T2 cases (observed in Table 9), is due to both false positive findings for the Tumour Size (TS) classifier, and the imperfect sensitivity and specificity of the Visceral Pleural Invasion (VPI) classifier. Erroneous T3 and T4 stage classifications are mostly due to the Chest Wall Invasion (CW) and the SEPN (Separate Tumour Nodules in Same Lobe) classifiers. The lower performance for those factors is consistent with both their rarity and the subjectivity seen in the corresponding expert decisions, as shown in Table 7.

Higher accuracy for N stage sentence-level factor results are likely to reflect the higher prevalence of N stage factors than T stage factors in the reports, however there is substantial agreement between the automatic classifiers and the experts for all N stage factors. As seen in the confusion matrix in Table 9, most of the system-level N stage errors are false positive findings of N0. These result from false negative findings from the

lymph node involvement classifiers (HLN, PLN, MLN and SLN) coupled with false positives from the NONM classifier.

B Other Considerations and Limitations

CSIS has been developed and evaluated for T and N staging of lung cancer based on reports from pathological studies of the lung, as proof-of-concept to determine the potential accuracy of an automatic system. There are several issues to be addressed for the system to be generalised to other cancers, or to process other input modalities before deployment in practice.

The current system was developed on a specific data set and there is a risk that over-fitting may limit broader application. Using more complex Natural Language Processing, richer medical terminologies (SNOMED CT, MetaMap), as well as larger and more varied training data sets may improve the generalisation and portability of classifiers to new cancers or reporting modalities.

A practical consideration is the expert time required for training SVM classifiers during system development. This involves annotation of sentences for each staging factor, which was done manually by the development team in the current system. It is estimated that the present lung cancer system involved up to 40 hours of annotation work during development. While this is not negligible, it must only be done once for each new cancer type, and is therefore not a major concern given eventual productivity gains from automatic stage data collection. Ongoing research is investigating methods for reducing annotation work in several ways, such as by identifying reusable classifiers across different cancers (e.g. tumour dimension, or the common ‘invasion’ classifier in the present system), analysing convergence with training set size, and using active learning.

Another practical consideration is the need to automatically discard irrelevant reports. The report relevance stage in the current system discards reports with no information for T or N staging, leading to TX or NX classifications. The system however assumes the input

reports do relate to lung cancer. This has been achieved in the development and trial data sets by filtering on report metadata (e.g. disease codes, examination type) from the source databases, however a practical system may require a more general report filtering stage. Much analysis of cancer outcomes is based on the higher level group stage, rather than the TNM stage. Because CSIS was developed on pathology reports and M staging is usually determined clinically or by medical imaging, M staging was omitted and CSIS therefore cannot output a proper group stage. Some indication of potential group stage accuracy can be given by assuming a known M stage. For all M0 cases with expert agreement on group stage from the trial, the present system attains an accuracy of 76.7% across Stages I-III (163 cases, Stage IV could not be assessed as it is defined as M1 with any T and N). Future work will investigate adaptability to using additional input sources, e.g. radiology or non-lung pathology reports, to determine M stage.

VI. CONCLUSIONS

We developed a prototype software system to automatically determine a patient's cancer stage from medical reports of lung cancer patients. The system uses Support Vector Machine classification techniques to classify a range of detailed staging factors at the sentence-level, and then combines these into a global stage decision. CSIS was compared against direct report-level classification and against staging by two independent pathology experts. The following conclusions can be made:

1. There is a significant level of disagreement in stage assigned by independent human experts based on pathology reports, particularly for T staging.
2. In comparison with human experts, CSIS achieved overall accuracy of 74% for T staging and 87% for N staging.
3. The two-level sentence classification approach improves on previous direct report-level stage classification by approximately 10% for both T and N staging.
4. The CSIS error pattern mirrors that observed between two independent experts.

The level of accuracy required for practical deployment of such a system would necessarily depend on the use case, and whether it involved a step of human validation. The results achieved do however lie within bounds of human staging accuracy observed in studies of registry data [4][5][6][7]. A productive avenue of research may be to improve the sensitivity of the N stage detail classifiers through more sophisticated natural language processing techniques. The limitations with the T staging system mostly reflect uncertainty in the report language, as well as the fact that the stage protocols do not cater for every contingency for more advanced cancer cases, thus leading to subjective interpretations. As well as investigating new classification strategies to improve sensitivity of detail classifiers, ongoing work will focus on addressing these issues for practical deployment of the technology.

VII. ACKNOWLEDGEMENTS

The research in this article was done in partnership with the Queensland Cancer Control and Analysis Team (QCCAT) within Queensland Health. In particular, the authors would like to acknowledge the following people: Hazel Harden, Shoni Colquist and Steven Armstrong from QCCAT; Jaccalyne Brady and Donna Fry from the Queensland Integrated Lung Cancer Outcomes Project; and Wayne Watson from the AUSLAB Support Group.

VIII. REFERENCES

[1] Clinical practice guidelines for the prevention, diagnosis and management of lung cancer. The Cancer Council Australia, 2004.

[2] F.L. Greene, D.L. Page, I.D. Fleming, A. Fritz, C.M. Balch, D.G. Haller, and M. Morrow, editors. AJCC Cancer Staging Manual. Springer, 6th edition, 2002.

[3] A.C. Klassen, F. Curriero, M. Kulldorff, A.J. Alberg, E.A. Platz, and S.T. Neloms. Missing stage and grade in Maryland prostate cancer surveillance data, 1992-1997. American Journal of Preventative Medicine, 30:577-587, 2006.

[4] J. Yau, A. Chan, T. Eapen, K. Oirourke, and L. Eapen. Accuracy of the oncology patients information system in a regional cancer centre. Oncology Reports, 9:167-169, 2002.

[5] W.K. Evans, J. Crook, D. Read, J. Morriss, and D.M. Logan. Capturing tumour stage in a cancer information database. Cancer Prevention and Control, 2:304-309, 1998.

[6] W-L. Liu, S. Kasl, J.T. Flannery, A. Lindo, and R. Dubrow. The accuracy of prostate cancer staging in a population-based tumor registry and its impact on the Black-White stage difference (Connecticut, US). Cancer Causes and Control, 6:425-430, 1995.

[7] L.J. Schouten, J.A. Langendijk, J.J. Jager, and P.A. van den Brandt. Validity of the stage of lung cancer in records of the Maastricht cancer registry, The Netherlands. Lung Cancer, 17:115-122, 1997.

[8] P. Silcocks, P. Needham, and F. Hemsley. Audit of prostate cancer: validity and feasibility of registry-based staging. Public Health, 113:157-160, 1999.

[9] T. Threlfall, J. Wittorff, P. Boutdara, L. Fritschi, J. Heyworth, P. Katris, and H. Sheiner. Collection of Population-based Cancer Staging Information in Western Australia - A Feasibility Study. Technical Report, National Cancer Control Initiative, Melbourne, 2004. Available at http://www.ncci.org.au/.

[10] P. Boutard, C. Platell, and T. Threlfall. Model for collecting colorectal cancer staging information in Western Australia. ANZ Journal of Surgery, 74:895-899, 2004.

[11] J.R. Condon, T. Barnes, B.K. Armstrong, S. Selva-Nayagam, and J.M. Elwood. Stage at diagnosis and cancer survival for indigenous Australians in the Northern Territory. Medical Journal of Australia, 182:277-280, 2005.

[12] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1–47, March 2002.

[13] I. McCowan, D. Moore, and M-J. Fry. Classification of cancer stage from free-text histology reports. In Proceedings of the IEEE Engineering in Medicine and Biology Conference, 2006.

[14] A. Nguyen, D. Moore, I. McCowan and M-J. Courage. Multi-class Classification of Cancer Stages from Free-text Histology Reports using Support Vector Machines. Submitted for publication, 2007.

[15] K. Aas and L. Eikvil. Text categorisation: A survey. Technical report, Norwegian Computing Center, 1999.

[16] V.N. Vapnik. The nature of statistical learning theory. Springer, 1995.

[17] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In Proc. 10th European Conference on Machine Learning, 1998.

[18] S.T. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In Proc. ACM-CIKM98, pages 148–155, 1998.

[19] D.D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. In Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.

[20] Y. Yang and X. Liu. A re-examination of text categorization methods. In Proceedings of 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999.

[21] G. Hripcsak, C. Friedman, P.O. Alderson, W. DuMouchel, S.B. Johnson and P.D. Clayton. Unlocking clinical data from narrative reports: A study of natural language processing. Annals of Internal Medicine, 122:681-688, 1995.

[22] W. Lehnert, S. Soderland, D. Aronow, F. Feng, and A. Shmueli. Inductive text classification for medical applications. Journal for Experimental and Theoretical Artificial Intelligence, 7(1):271–302, 1995.

[23] D.B. Aronow, F. Fangfang, and W.B. Croft. Ad hoc classification of radiology reports. Journal of the American Medical Informatics Association, 6(5):393–411, 1999.

[24] Y. Aphinyanaphongs, I. Tsamardinos, A. Statnikov, D. Hardin, and C.F. Aliferis. Text categorization models for high-quality article retrieval in internal medicine. Journal of the American Medical Informatics Association, 12(2):207–216, 2005.

[25] W. Chapman, L. Christensen, M. Wagnera, P. Haug, O Ivanov, J. Dowling, and R. Olszewski. Classifying free-text triage chief complaints into syndromic categories with natural language processing. Artificial Intelligence in Medicine, 33(1):31–40, 2004.

[26] W.W. Chapman, J.N. Dowling, and M.M. Wagner. Fever detection from free-text clinical records for biosurveillance. Journal of Biomedical Informatics, 37:120–127, 2004.

[27] W.W. Chapman, G.F. Cooper, P. Hanbury, B.E. Chapman, L.H. Harrison, and M.M. Wagner. Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders. Journal of the American Medical Informatics Association, 10:494–503, 2003.

[28] B. de Bruijn, A. Cranney, S. O'Donnell, J.D. Martin, and A.J. Forster. Identifying wrist fracture patients with high accuracy by automatic categorization of x-ray reports. Journal of the American Medical Informatics Association, 13:696-698, 2006.

[29] O. Ivanov, M.M. Wagner, W.W. Chapman, and R.T. Olszewski. Accuracy of three classifiers of acute gastrointestinal syndrome for syndromic surveillance. In Proceedings of the American Medical Informatics Association Symposium, 2002.

[30] Q. Zeng, S.Goryachev, S.Weiss, M.Sordo, S. Murphy and R.Lazarus. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Medical Informatics and Decision Making, 6, 2006.

[31] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics, 34:301–310, 2001.

[32] W.W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, and B.G. Buchanan. Evaluation of negation phrases in narrative clinical reports. In Proceedings of American Medical Informatics Association Symposium, 2001.

[33] W.W. Chapman, M. Fizman, B.E. Chapman, and P.J. Huang. A comparison of classification algorithms to automatically identify chest xray reports that support pneumonia. Journal of Biomedical Informatics, 34:4–14, 2001.

[34] M. Sordo and Q. Zeng. On sample size and classification accuracy: a performance comparison. Lecture Notes in Computer Science, 3745:193-201, 2005.

[35] PlatoCode. http://www.platocode.com/, 2006.

[36] A-Life Medical. http://www.alifemedical.com/, 2006.

[37] AIM. http://www.aim.on.ca/, 2006.

[38] P. Phinjaroenphan and S. Bevinakoppa. Automated prognostic tool for cervical cancer patient database. In Proceedings of International Conference on Intelligent Sensing and Information Processing, 2004.

[39] P. Mitra, S. Mitra, and S.K. Pal. Staging of cervical cancer with soft computing. IEEE Transactions on Biomedical Engineering, 47(7):934–940, 2000.

[40] C. Renz, J.C. Rajapakse, K. Razvi, and S.K.C. Liang. Ovarian cancer classification with missing data. In Proceedings of the 9th International Conference on Neural Information Processing, 2002.

[41] A. Tewari and P. Narayan. Novel staging tool for localized prostate cancer: a pilot study using genetic adaptive neural networks. Journal of Urology, 160:430-436, 1998.

[42] M. Han, P.B. Snow, J.M. Brandt, and A.W. Partin. Evaluation of artificial neural networks for the prediction of pathologic stage in prostate carcinoma. Cancer, 91:1661-1666, 2001.

[43] E.J. Gamito, N.N. Stone, J.T. Batuello, and E.D. Crawford. Use of artifical neural networks in the clinical staging of prostate cancer: implications for prostate brachytherapy. Techniques in Urology, 6:60-63, 2000.

[44] mTuitive. http://www.mtuitive.com/, 2006.

[45] Collaborative Staging. http://www.cancerstaging.org/cstage/index.html, 2006.

[46] J. Kupiec, J. Pedersen and F. Chen. A Trainable Document Summarizer, In Proceedings of 18th ACM SIGIR Conference, 68-73, 1995.

[47] S. Teufel and M. Moens. Sentence Extraction as a Classification Task, In Proc. ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, 58-68, 1997.

[48] L. McKnight and P. Srinivasan. Categorization of Sentence Types in Medical Abstracts, In Proceedings American Medical Informatics Association Symposium, 440-444, 2003.

[49] A. De Sitter and W. Daelemans. Information Extraction via Double Classification. In Proc. International Workshop on Adaptive Text Extraction and Mining, 2003.

[50] NIH. Unified Medical Language System (UMLS), 2006. http://www.nlm.nih.gov/research/umls/.

[51] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART: TREC 3. In Proceedings of the 3rd Text Retrieval Conference. NIST, 1994.

[52] T. Joachims. Making large-scale SVM learning practical. In B. Scholkopf, C. Burges, and A. Smola, ed, Advances in Kernel Methods - Support Vector Learning. MIT Press, 1999.

[53] K.M. Fong, R. Bowman, D. Fielding, R. Abraham, M. Windsor M, and G. Pratt. Queensland integrated lung cancer outcomes project (QILCOP): Initial accrual and preliminary data from the first 30 months. Abstract of Presentation at The Thoracic Society of Australia and New Zealand Annual Scientific Meeting, April 2003.

[54] D. Moore, I. McCowan, A. Nguyen and M-J Fry. Trial evaluation of automatic lung cancer staging from pathology reports. In Proceedings of 12th International Health (Medical) Informatics Congress (Medinfo), 2007.