

# Automated Cancer Stage Classification from Free-text Histology Reports

Iain McCowan<sup>1</sup>, Darren Moore<sup>1</sup>, Mary-Jane Fry<sup>2</sup>

<sup>1</sup>CSIRO E-Health Research Centre, Brisbane, Australia.

<sup>2</sup>Queensland Cancer Control Analysis Team (QCCAT), Queensland Health, Brisbane, Australia.

## ABSTRACT

**Objectives:** This article describes a system to automatically classify the stage of a lung cancer patient based on text analysis of their histology reports. **Methods:** The system uses machine learning techniques to train a statistical classifier, specifically a support vector machine, for each TNM stage category based on word occurrences in a corpus of histology reports for staged patients. New reports can then be classified according to the most likely stage, facilitating the collection and analysis of population staging data. While the system could in principle be applied to stage different cancer types, the current work focuses on staging lung cancer due to data availability. **Results:** The article presents initial experiments quantifying system performance on a corpus of reports from more than 1000 lung cancer patients. Results give average sensitivity of 0.72 and specificity of 0.87 for pathologic staging based on histology report text.

## Keywords:

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

## 1. INTRODUCTION

The *cancer stage* categorises a cancer's progression in the body, in terms of the extent of the primary tumour and any spreading to local or distant body sites. Routine staging of cancer patients has a number of benefits recognised by cancer bodies worldwide: it allows a physician to determine treatment more appropriately, evaluate outcomes more reliably, and compare statistics on a local, regional, and national basis more confidently. These benefits have motivated international standards for cancer staging, including 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 [1]. Routine staging of patients according to this system is increasingly being recommended as a standard of care by national cancer bodies, e.g. [2].

In spite of its recognised utility, formal staging data is not routinely collected for all cancer patients. For instance, according to [3] in 2004 there was no on-going population-based collection of staging information in any Australian state or territory. Efforts since have aimed to rectify this, however it is not expected that current methods for cancer staging will be applied for all patients due to the time- and resource-intensive nature of multi-disciplinary team conferences (MDC's). Technological support for the cancer stage decision has been limited to date. While some software products exist to assign a TNM stage (e.g. [4, 5]), these generally rely on highly structured input, and therefore do not reduce the need for expert reading and interpretation of reports, or require significant changes to reporting systems and practice.

This article describes initial work towards a decision support system for staging cancer patients based on free-text medical reports. While the system could in principle be applied to stage other cancers, the present article focuses on staging lung cancer for reasons of data availability. For a given patient, the input to the system consists of a variable number of textual reports describing the results of histology tests. The objective of the system is to determine T, N, and M stage values for the patient. The system aims to achieve this by applying machine learning *text categorisation* techniques [6].

Text categorisation (see [6, 7] for recent reviews) is the task of deciding if a document belongs to each of a set of predefined categories. Very early work in this field focussed on knowledge-based approaches, mainly consisting of manual definition of sets of rules that attempt to encode the expert knowledge required to categorise documents. The major disadvantage with these approaches is the need for human experts to define

<b>T: Primary Tumour</b>	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.
<b>N: Regional Lymph Nodes</b>	X	Regional lymph nodes cannot be assessed.
	0	No regional lymph node metastasis.
	1,2,3	Increasing involvement of regional lymph nodes.
<b>M: Distant Metastasis</b>	X	Distant metastasis cannot be assessed.
	0	No distant metastasis.
	1	Distant metastasis.

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

and maintain the comprehensive rule set required for high accuracy. For this reason most text categorisation research in recent years has concentrated on machine learning approaches which automatically build text classifiers by learning the characteristics of each category from a set of pre-classified documents (the *training corpus*). Such a machine learning approach is taken in the present system.

Most medical-related automated text analysis work in the literature has dealt with the problem of converting free-text reports into standard codes or structured formats that are more suitable for further analysis, e.g. [8, 9, 10]. Beyond such automatic coding systems, there have been a number of systems that have attempted classification of medical reports, for instance according to specific medical diseases or conditions. This has included: classification of radiology reports according to 6 conditions [11], customisable decision-tree inductive classifiers [12, 13], classification of high quality MEDLINE articles [14], classification of emergency department reports into eight syndromic categories [15], detecting fever in emergency department patients [16], detection of radiology reports that support a finding of inhalational anthrax [17], and detection of acute gastrointestinal syndrome of public health significance from emergency department reports [18].

Literature and market reviews have uncovered few instances of research or commercial software systems that specifically assist in the cancer stage decision. The stage of cervical cancer was determined by a neural network classifier in [19], using a 15-element vector encoding high level results of MRI and PET scans as input. A soft-computing approach was used in [20] to classify cervical cancer cases into one of 4 FIGO stages based on a vector encoding the presence or absence of each major symptom. The mTuitive [4] xPert product line includes a module for cancer staging according to the AJCC TNM guidelines, based on structured data entry. The Collaborative Staging Task Force [5] has produced a set of common software tools to determine the cancer stage according to multiple systems. The system takes as input a structured set of all data items required for a given cancer type and then applies a deterministic algorithm to assign the correct stage code. The system proposed in the current article can be clearly differentiated from the above systems for cancer staging in two main ways: firstly in its use of free-text reports rather than highly structured input data, and secondly as it uses probabilistic rather than deterministic algorithms - this may be important when only partial and uncertain information is available, such as during initial clinical staging, and also when access to expert knowledge of staging is limited.

The remainder of this article is organised as follows. Section 2 describes the proposed method for classifying the cancer stage. An experimental evaluation of the system is presented in Section 3 followed by ongoing work and concluding remarks in Section 4.

## 2. METHOD

For each patient, the input to the system consists of unstructured text taken from available histology reports. The text is first normalised to reduce basic variations: the formats of acronyms, numbers and dimensions are standardised, relevant abbreviations are expanded, spelling variants are mapped to a common form, and any non-informative character sequences are removed. The set of normalisation rules are encoded using regular expressions and implemented using simple search and replace operations.

Following normalisation, the text is converted into a sequence of base forms from the Unified Medical Language System (UMLS) SPECIALIST Lexicon [21]. The SPECIALIST Lexicon is a general English lexicon supplemented with many single or multiple word biomedical terms. Each lexical record has a base form, which

is the uninflected form of the term (e.g. singular form for nouns, infinitive form for verbs). The utility of the UMLS SPECIALIST Lexicon has been previously demonstrated for similar tasks, e.g. in a noun phrase identification system for radiology reports in [22]. The conversion to UMLS base form yields the dual benefits of adhering to an open standard lexicon, and effectively implementing a stemming step to further reduce variability of the report text.

Following these text pre-processing steps, features are extracted for classification. A vector space model is used to represent each text report as a vector of term (word) weights. The term weights are calculated according to the LTC-weighting scheme [23, 7], which is commonly used in state-of-the-art text categorisation systems, as it effectively de-emphasises common terms (occurring often in many reports), produces normalised weights across different length reports, and reduces the impact of large differences in frequency.

The final step in the system is to classify the cancer stage from the LTC features using Support Vector Machines (SVM's) [24, 25]. For each of the cancer stage categories (from Table 1, e.g. T1, N2), a binary SVM classifier is trained based on whether each report in a training corpus is relevant to that particular category. The SVM's are implemented using the open source SVM<sup>light</sup> toolkit [26]. The parameters of the SVM are estimated from a training corpus of text reports supplemented with MDC stage data. During testing, the SVM outputs a score that can be thresholded to decide if a new document belongs to a particular class. Classifiers are not trained for the TX, NX and MX stage categories, as these may be considered as defaults if no other categories can be assigned.

### 3. RESULTS AND DISCUSSION

To train and validate the system, a corpus of de-identified medical reports with corresponding staging data was obtained for 1054 lung cancer patients following research ethics approval. The corpus was compiled from two separate data sources: a database of pathologic staging decisions for lung cancer patients (Queensland Integrated Lung Cancer Outcomes Project data [27]) for use as ground-truth for the classifier training and testing, and a set of histology reports for lung cancer patients extracted from the state pathology information system (AUSLAB). In order to maximise the amount of SVM training data while still reporting significant results on this dataset, an N-fold scheme was applied. The data was randomly divided into 100 subsets (approximately 10 patients per subset). In each fold, system output was generated for one subset from an SVM trained on the remaining 99 subsets. This meant that over the 100 folds, results could be reported on the full 1054 patients while ensuring each result was produced by an unbiased system (where test data was not used during system training).

In medical literature the most common measures for binary tests are sensitivity, specificity and positive predictive value (PPV). Given the True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN), then  $Sensitivity = TP / (TP + FN)$ ,  $Specificity = TN / (TN + FP)$ , and  $PPV = TP / (TP + FP)$ . In the text classification literature, the most common measures are recall and precision. In the current context, *recall* is the same as sensitivity, while *precision* is the same as PPV. If a single performance measure is required, the *F1-measure* is commonly used in the text classification literature; this is the harmonic mean of precision and recall. A more naïve measure of system performance is given by the *accuracy*, which is the proportion of reports that were correctly assigned by that classifier (true positive and true negatives). Each of these values can first be calculated on a per-category basis and then averaged *across categories* to give *macro-averaged* results, or averaged *across all patients* to give micro-averaged results. In general micro-averaged scores tend to be dominated by classifier performance on the most common categories, while macro-averaged are influenced more by classifier performance on rare categories. Depending on the application, a trade-off exists between sensitivity and specificity, or recall and precision, and this can be controlled by varying one or more classifier hyper-parameters. It is common to report performance at a task-relevant operating point, such as the *break even point* (where two complementary evaluation measures are identical).

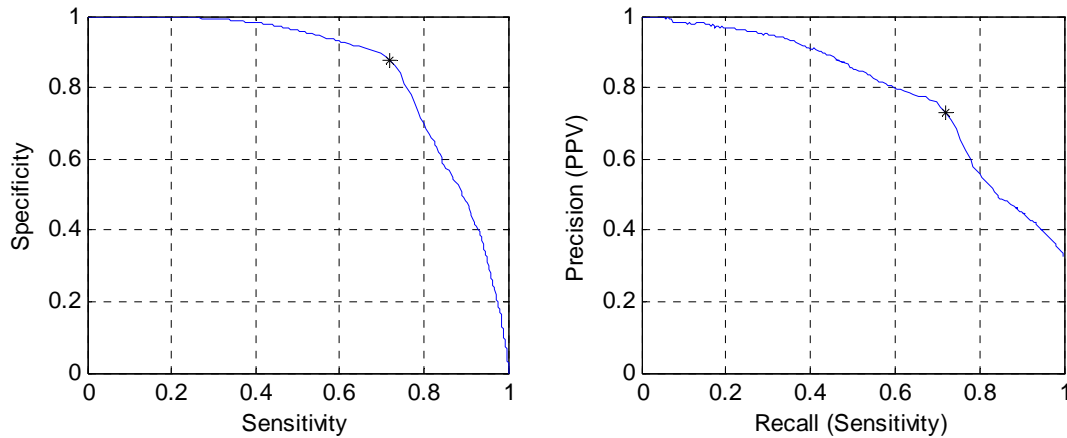
Table 2 presents these performance measures for each binary SVM classifier (i.e. per stage category); results are not presented for T0, Tis, and N3 as these categories had insufficient positive examples for classifier training ( $P < 100$ ). Results are reported at the precision (PPV) / recall (sensitivity) break even point. The results show some interesting trends: specificity is consistently high across categories indicating strong reliability in a negative decision from each classifier; sensitivity and PPV (recall and precision) show that the classifiers perform significantly above chance levels in their positive stage decisions; and in general classifiers with significant

training data (both high P and N) perform better, achieving above 0.64 across all measures (e.g. N0, N1, T1, T2). This last point indicates that overall performance may be expected to improve significantly with a larger development dataset. In terms of overall system performance, the macro- and micro-averaged F1 measures are 0.621 and 0.725 respectively, which is a promising result for these preliminary experiments. Accuracy figures are less informative, due to the unbalanced number of positive and negative examples for most classifiers.

Stage	P	N	Specificity	Sensitivity	PPV	F1	Accuracy
M0	900	154	0.461	0.900	0.907	0.904	0.836
M1	151	903	0.914	0.477	0.480	0.478	0.851
N0	588	466	0.717	0.767	0.774	0.770	0.745
N1	180	874	0.943	0.644	0.699	0.671	0.892
N2	178	876	0.894	0.483	0.480	0.482	0.824
T1	258	796	0.899	0.647	0.676	0.661	0.838
T2	530	524	0.716	0.692	0.711	0.702	0.704
T3	123	931	0.921	0.472	0.439	0.455	0.868
T4	132	922	0.924	0.462	0.466	0.464	0.866
<b>Macro-average</b>			0.821	0.616	0.626	<b>0.621</b>	0.825
<b>Micro-average</b>			0.874	0.720	0.730	<b>0.725</b>	0.825

**Table 2: Results for each stage classifier with macro- and micro-averages: Positive cases (P), Negative cases (N), Sensitivity (Recall), Specificity, Positive Predictive Value (PPV, Precision), F1 Measure and Accuracy.**

As mentioned previously, there is a trade-off between system performance for complementary measures. Depending on the relative application-dependent “costs” (financial, emotional or otherwise) a false negative finding may be favoured over a false positive finding, or vice-versa. Figure 1 plots the receiver operator characteristic (ROC) curves for complementary micro-averaged results by varying the SVM decision threshold. This shows, for example, that average sensitivity greater than 0.9 may be achieved at the cost of specificity of 0.5 or below. Another useful single performance measure is the area under the ROC curve; from Figure 1, this is 0.851 for sensitivity-specificity, and 0.790 for recall-precision, which are again promising initial results.



**Figure 1: Receiver Operator Characteristic (ROC) Curves for Specificity vs Sensitivity (area = 0.851), and Precision (PPV) vs Recall (Sensitivity) (area = 0.790). The operating point for Table 2 results is marked by \*.**

#### 4. CONCLUSION

This article has presented preliminary progress towards a system to assist in the collection of staging data for lung cancer patients. Initial results show a system based on standard support vector machine classifiers achieves average sensitivity of 0.72 and specificity of 0.87 for pathologic staging based on histology report text. While this is clearly promising, there is much scope to improve the system to incorporate specific knowledge of the staging protocol. Ongoing work will investigate the enhancement of the system with natural language processing techniques, e.g. to detect negated findings (“no evidence of pleural invasion”), as well as rules specific to each stage category, e.g. associating dimensions with key terms (“primary tumour greater than 3cm in extent”). The system will also be extended to use radiology reports and to support clinical staging.

## ACKNOWLEDGEMENTS

This research was done in partnership with the Queensland Cancer Control and Analysis Team (QCCAT). In particular, the authors wish to acknowledge: Hazel Harden, Shoni Colquist and Steven Armstrong from QCCAT for their help in system concept definition, and for providing access to data and clinical experts; Jaccalyne Brady and Donna Fry from QILCOP for explaining cancer stage decision making processes; Dr Rayleen Bowman and Dr Belinda Clarke for their expert advice; and Wayne Watson from the AUSLAB Support Group for extracting histology reports to form the research data corpus.

## REFERENCES

- [1]. Greene, F.; Page, D.; Fleming, I.; Fritz, A.; Balch, C.; Haller, D. & Morrow, M., ed., *AJCC Cancer Staging Manual*, Springer, 2002.
- [2]. *Clinical practice guidelines for the prevention, diagnosis and management of lung cancer*, The Cancer Council Australia, 2004.
- [3]. Threlfall, T. et al. Collection of Population-based Cancer Staging Information in Western Australia - A Feasibility Study, National Cancer Control Initiative, 2004.
- [4]. mTuitive, <http://www.mtuitive.com/>
- [5]. Collaborative Staging, <http://www.cancerstaging.org/cstage/index.html>
- [6]. F. Sebastiani. Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, 2002, 34, 1-47.
- [7]. Aas, K. & Eikvil, L. Text categorisation: A survey Norwegian Computing Center, 1999.
- [8]. Cooper, G. & Miller, R. An experiment comparing lexical and statistical methods for extracting MeSH terms from clinical free text, *Journal of the American Medical Informatics Association*, 1998, 5, 62-75.
- [9]. Hazlehurst, B.; Frost, H.; Sittig, D. & Stevens, V. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record, *Journal of the American Medical Informatics Association*, 2005, 12, 517-529.
- [10]. Dolin, R.; Aschuler, L.; Beebe, C.; Biron, P.; Boyer, S. & Essin, D. The HL7 Clinical Document Architecture, *Journal of the American Medical Informatics Association*, 2001, 8, 552-569.
- [11]. Wilcox, A. & Hripcsak, G. The role of domain knowledge in automating medical text report classification, *Journal of the American Medical Informatics Association*, 2003, 10, 330-338.
- [12]. Lehnert, W.; Soderland, S.; Aronow, D.; Feng, F. & Shmueli, A. Inductive Text Classification for Medical Applications, *Journal for Experimental and Theoretical Artificial Intelligence*, 1995, 7, 271-302.
- [13]. Aronow, D.; Fangfang, F. & Croft, W. Ad hoc classification of radiology reports *Journal of the American Medical Informatics Association*, 1999, 6, 393-411.
- [14]. Aphinyanaphongs, Y. et al. Text categorization models for high-quality article retrieval in internal medicine, *Journal of the American Medical Informatics Association*, 2005, 12, 207-216.
- [15]. Chapman, W.; Christensen, L.; Wagner, M.; Haug, P.; Ivanov, O.; Dowling, J. & Olszewski, R. Classifying free-text triage chief complaints into syndromic categories with natural language processing, *Artificial Intelligence in Medicine*, 2004, 33, 31-40.
- [16]. Chapman, W.; Dowling, J. & Wagner, M. Fever detection from free-text clinical records for biosurveillance, *Journal of Biomedical Informatics*, 2004, 37, 120-127.
- [17]. Chapman, W.; Cooper, G.; Hanbury, P.; Chapman, B.; Harrison, L. & Wagner, M. 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*, 2003, 10, 494-503.
- [18]. Ivanov, O.; Wagner, M.; Chapman, W. & Olszewski, R. Accuracy of three classifiers of acute gastrointestinal syndrome for syndromic surveillance. *Proc American Medical Informatics Association Symp.* 2002:345-9. 2002.
- [19]. P.Phinjaroenphan & S.Bevinakoppa. Automated prognostic tool for cervical cancer patient database 2004.
- [20]. Mitra, P.; Mitra, S. & Pal, S. Staging of cervical cancer with soft computing *IEEE Transactions on Biomedical Engineering*, 2000, 47, 934-940.
- [21]. Unified Medical Language System, <http://www.nlm.nih.gov/research/umls/>
- [22]. Huang, Y.; Lowe, H.; Klein, D. & Cucina, R. Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the UMLS specialist lexicon, *Journal of the American Medical Informatics Association*, 2005, 12, 275-285.
- [23]. Buckley, C.; Salton, G.; Allan, J. & Singhal, A. Automatic Query Expansion Using SMART: TREC 3 NIST, 1994.
- [24]. Vapnik, V. *The nature of statistical learning theory*, Springer, 1995.
- [25]. Joachims, *Text categorization with support vector machines: Learning with many relevant features*, 1998.
- [26]. Joachims, T. Schölkopf, B.; Burges, C. & Smola, A. (ed.) *Making large-scale SVM learning practical*, MIT Press, 1999.
- [27]. Fong, K.; Bowman, R.; Fielding, D.; Abraham, R.; M, M.W. & Pratt, G. Queensland Integrated Lung Cancer Outcomes Project (QILCOP): Initial Accrual and Preliminary Data from the First 30 Months, 2003.

### Address for Correspondence

Iain McCowan, EHRC, PO Box 10842 Adelaide Street, Brisbane Q 4000. [iain.mccowan@csiro.au](mailto:iain.mccowan@csiro.au)