Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans

David Martinez a, Michelle R. Ananda-Rajah b, Hanna Suominen c,d,e, Monica A. Slavin f,g, Karin A. Thursky f,g, Lawrence Cavedon h,⇑

a CIS Department, University of Melbourne, Australia
b Infectious Diseases Unit, Alfred Health and The University of Melbourne, Australia
c NICTA and The Australian National University, Canberra, Australia
d University of Canberra, Canberra, Australia
e University of Turku, Finland
f Victorian Infectious Diseases Service, Royal Melbourne Hospital, Peter MacCallum Cancer Institute, Australia
g Infectious Diseases Department, Peter MacCallum Cancer Institute, Australia
h School of Computer Science and IT, RMIT University, Melbourne, Australia

ABSTRACT

Background: Invasive fungal diseases (IFDs) are associated with considerable health and economic costs. Surveillance of the more diagnostically challenging invasive fungal diseases, specifically of the sino-pulmonary system, is not feasible for many hospitals because case finding is a costly and labour intensive exercise. We developed text classifiers for detecting such IFDs from free-text radiology (CT) reports, using machine-learning techniques.

Method: We obtained free-text reports of CT scans performed over a specific hospitalisation period (2003–2011), for 264 IFD and 289 control patients from three tertiary hospitals. We analysed IFD evidence at patient, report, and sentence levels. Three infectious disease experts annotated the reports of 73 IFD-positive patients for language suggestive of IFD at sentence level, and graded the sentences as to whether they suggested or excluded the presence of IFD. Reliable agreement between annotators was obtained and this was used as training data for our classifiers. We tested a variety of Machine Learning (ML), rule based, and hybrid systems, with feature types including bags of words, bags of phrases, and bags of concepts, as well as report-level structured features. Evaluation was carried out over a robust framework with separate Development and Held-Out datasets.

Results: The best systems (using Support Vector Machines) achieved very high recall at report- and patient-levels over unseen data: 95% and 100% respectively. Precision at report-level over held-out data was 71%; however, most of the associated false-positive reports (53%) belonged to patients who had a previous positive report appropriately flagged by the classifier, reducing negative impact in practice.

Conclusions: Our machine learning application holds the potential for developing systematic IFD surveillance systems for hospital populations.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Invasive fungal diseases (IFDs) are associated with considerable health [27,38] and economic costs [2,32]. Among the immunocompromised host population, patients with haematological malignancies and haematopoietic stem cell transplant recipients (HSCTs) carry the greatest burden of IFDs [27,38]. The incidence of IFDs is highly variable, ranging from 8% to 48% in patients with acute myeloid leukaemia and HSCT patients [27,35]. IFDs of the sino-pulmonary system, of which invasive aspergillosis (IA) is most common, now comprise the majority of IFDs [38,27]—focus in this paper is exclusively on this class of diagnostically challenging IFDs, including invasive aspergillosis. The short-term mortality of invasive pulmonary aspergillosis remains unacceptably high at 34–43% in more recent reports among HSCT recipients and patients with a variety of haematological malignancies [37,35,42]. Given the high health and economic burden of IFDs, several authorities and professional
 societies advocate that surveillance of IFDs should be the standard of care [22,45,53]. Surveillance is a necessary step towards defining the burden of IFDs, informing and evaluating choice of preventative strategies, tracking epidemiological data in response to changing therapeutic advances, host or environmental factors [39], and recognising sporadic but catastrophic healthcare related outbreaks [26]. At present, surveillance of IFDs is not routinely performed in most hospitals for a variety of reasons, including cost and the absence of an easily identifiable laboratory prompt (such as a positive blood culture) for mould diseases. Laboratory-based surveillance for IFDs is not suitable because isolation of infection-causing moulds occurs in less than 50% of cases [15] and/or patients may be too unwell to undergo invasive diagnostic procedures. Further, non-culture based tests are not widely available in hospitals, or may be associated with a delayed turn-around. Clinical review is challenging, requiring multiple data sources (radiology, clinical, laboratory) and multi-disciplinary teams, followed by the application of complicated definitions [13,27,36]. Thus, traditional methods of surveillance using either clinical review, laboratory-based methods, or less commonly administrative data are costly, labour intensive, error prone, and subject to incomplete case findings [9,27].

The optimal screening method for IFD surveillance is undefined but the choice of screening method is critically important to minimise effort while maximising case capture. Computed tomography (CT) is appealing as a screening method for IFD surveillance. CT is a key diagnostic test for IFDs stipulated in internationally recognised guidelines [13], and lung involvement is present in the overwhelming majority (90–100%) of patients with IFDs [27,38,36]. CT is a non-invasive test uniformly performed when IFD is suspected and it is widely available in hospitals with results reported within hours rather than days. Although the radiologic features of IFDs are not specific for IFD [29], CT remains a valuable diagnostic adjunct [13].

In this paper, we address the challenge of surveillance of sino-pulmonary IFDs using the technique of directly processing free-text radiology reports, specifically CT scans, as a means of screening patients for features supportive of IFD. Text Mining techniques over scans have been previously proposed to support detection and surveillance, including for infectious diseases and their symptoms (e.g., fever) [11,46,17,41,34,25,30], but not previously to the challenge of detecting IFDs. The approach has the potential to identify patients with suspected IFDs in real-time, delivering to hospitals a feasible, sustainable and cost-effective solution to the task of IFD surveillance with minimal interruption to routine clinical workflow.

The described approach is a novel methodological contribution to the task of identifying cases of a specific condition from CT scan reports. Text classifiers were developed for CT scan reports at the sentence, individual report, and patient level. Classification at the report level enables the potential for real-time detection and monitoring of incidence of IFDs, while patient-level classification enables a surveillance and reporting mechanism for IFD. Sentence-level classification is shown to facilitate improved classification performance at the report- and patient-levels, and has the added benefit of indicating “supporting evidence” for a positively-classified report. The collection of scan reports used was drawn from three major Melbourne hospitals. A subset of these reports was manually annotated by infectious disease physicians with domain expertise in the area of IFDs. Manual annotation was performed at the sentence level, where the physicians arbitrated over several categories according to pre-specified annotation guidelines, of which presence or absence of language supportive of the presence of an IFD was most discriminatory. Note that we use a light annotation scheme, requiring much less annotation effort than is generally required for machine-learning approaches to biomedical text mining tasks (which may require annotation of phrases and relationships between entities). Moreover, we ensured that the machine-learning classifiers have no hand-coded intervention: i.e., they were constructed in a fully automated fashion. This makes the techniques more transferable to other problem tasks, and means even better performance could potentially be achieved on deployment in specific contexts, via hand-crafted tuning.

2. Related work

As noted by Demner-Fushman et al. [14], radiology reports are a rich source of knowledge and were used in early applications of still-influential clinical NLP systems (e.g., [24,20]). The types of reports examined have varied (e.g., X-ray, CT), and tasks have ranged from specific classification tasks to more general named entity recognition (e.g., [52]), coding, and information extraction (e.g., [18,19]), across a broad range of tasks. Applications have included disease/infection detection and surveillance from radiology reports, with pneumonia being the disease of greatest focus. A number of authors have demonstrated favourable performance of more sophisticated NLP techniques for identifying/classifying radiology reports for specific purposes, in comparison to simpler techniques. For example, Solti et al. [40] demonstrated the efficacy of machine-learning (ML) NLP over keyword search for identifying X-ray reports of cases of acute lung injury; Womack et al. [51] showed that the NegEx system [10] compares favourably to keyword search for identifying cases of acute fracture. Table 1 summarises the main works discussed in this section.

Work associated with LDS Hospital from Salt Lake City (Utah, US) has resulted in a series of systems for coding concepts and performing disease and infection surveillance, particularly the identification of pneumonia cases from X-ray reports. Haug et al. [24] used Augmented Transition Networks (ATNs) for syntactic-level processing of X-ray reports and a Bayesian Network model for semantic-level representation and constraining the interpretation of concept terms. The system was evaluated on the task of identifying (three) pneumonia-related concepts and inferring the presence (or absence) of acute bacterial pneumonia from X-ray reports, and was found to perform to a level comparable to physicians: 0.95/0.78/0.85 recall/precision/specificity respectively for the system vs. 0.94/0.87/0.91 for the team of three physicians employing majority vote.

More recently, Tinoco et al. [43] compared the effectiveness of a more general version of LDS Hospital’s computerised surveillance system (CSS) against manual methods (i.e., chart review) on the tasks of detecting reportable adverse drug effects (ADE) and hospital-acquired infections (HAIs). They found that the automated system detected substantially more HAIs than did manual review (92% vs. 34%) while the two methods did not differ substantially on detecting ADEs, demonstrating the value of automated surveillance techniques for HAIs in general. One significant source of HAIs missed by the CSS was information in physician’s narratives—58% of the HAIs missed by the CSS were explicitly described in the narratives: i.e., adding text mining capabilities would have improved performance of HAI-detection up to approximately 97%.

The general clinical NLP system MEDLEE has been used for a number of tasks involving processing radiology reports, including in early work by its authors [20]. Of specific relevance to our context, Mendonça et al. [31] described the application of MEDLEE to
the detection of pneumonia in newborns, from a range of test sources, including radiology, pathology and microbiology reports. They demonstrated strong performance, compared to judgements made by clinicians: 71% sensitivity and 95% specificity.

Travers et al. [46] described a search system for identifying patients with acute respiratory illness, such as influenza, SARS, and anthrax. The system pre-processed text (e.g., replacing acronyms, abbreviations, and common misspellings), and clinician-designed queries were used to identify records indicating such cases. Specificity was high (0.99) but sensitivity was low (0.24).

Jones et al. [25] described a surveillance approach using a simple NLP pipeline to detect methicillin-resistant Staphylococcus aureus within US Veterans’ Affairs (VA) medical centres. Their NLP pipeline mapped concepts to SNOMED-CT and used rules based on matching against relevant keywords. With the benefit of a large data-collection (more than 65,000 records) available for training, they obtained 99% sensitivity and specificity. Matheny et al. [30] also tackled the task of detecting infectious symptoms from VA patient records. Their NLP approach also extracted concepts (against UMLS, HL-7, and SNOMED-CT) and also used rules based on concept- and keyword-matching.

Elkin et al. [17] used an NLP pipeline to identify cases of pneumonia, and related pulmonary diseases, by processing radiology reports (chest x-rays and CT scans), in a task closely related to ours. Their rule-based approach, however, was very different to our machine-learning approach. Elkin et al. used an NLP pipeline that identified clinical concepts in the text reports and mapped them into SNOMED-CT concepts. They then hand-authored rules operating over the SNOMED-CT concepts to attempt to identify positive (or uncertain) assertions of pneumonia (or positive assertions of infiltrates or consolidations when pneumonia assertions are not identified) in various sections of each radiology report. The rule-set for classifying “uncertain” cases (as well as “positive”) led to high performance: 1.0 sensitivity and 0.98 specificity. The risk with their approach is whether the rules were overfit to the data, since only 400 reports were used and testing occurs over the same dataset used to construct the rules.

Bejan et al. [7] tackled the problem of identifying intensive care patients with pneumonia from narrative reports, using techniques with some commonalities to those we describe below. In particular, they combined the use of n-grams with use of UMLS concepts as features. They used a feature-selection technique based on statistical significance testing to rank features, which resulted in large performance improvements over their baseline.

Finally, another recent work explored the detection of HAIs from text, in this case over Swedish patient records [16]. The annotation for this work was performed at patient level over 213 records, and also aimed at achieving high recall. They obtained close to 90% recall for 67% precision at patient level, and highlighted feature selection as a crucial step for raising recall.

As clear from the summary in Table 1, much of the previous related work used grammatical rule-based approaches for processing reports, the most sophisticated of which mapped concepts to a semantic representation language (e.g., as for MEDLEE) or standard ontologies (such as SNOMED-CT). While our approach has some similarities to the more recent approaches that used machine-learning techniques, an important point of difference is the hierarchical approach we employ, whereby sentence-level classification is used as a feature for report-level classification.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Approach</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solti et al. [40]</td>
<td>ML-based NLP vs. keyword-search</td>
<td>Identifying lung injury from X-ray reports</td>
</tr>
<tr>
<td>Haug et al. [24]</td>
<td>Rule-based NLP (ATNs)</td>
<td>Processing X-ray reports</td>
</tr>
<tr>
<td>Tinoco et al. [43]</td>
<td>Rules-based NLP (ATNs)</td>
<td>Identifying ADEs and HAIs</td>
</tr>
<tr>
<td>Mendonça et al. [31]</td>
<td>Rule-based NLP (MEDLEE)</td>
<td>Detecting pneumonia from different sources</td>
</tr>
<tr>
<td>Travers et al. [46]</td>
<td>Manually-constructed queries</td>
<td>Identifying patients from respiratory illness</td>
</tr>
<tr>
<td>Jones et al. [25]</td>
<td>Rule-based NLP, concept-mapping to SNOMED-CT</td>
<td>Detecting Staphylococcus aureus from VA records</td>
</tr>
<tr>
<td>Matheny et al. [30]</td>
<td>Concept/keyword-matching rules</td>
<td>Detection infection symptoms from VA records</td>
</tr>
<tr>
<td>Elkin et al. [17]</td>
<td>Rule-based NLP, concept-mapping to SNOMED-CT</td>
<td>Detecting pneumonia from X-rays/CT-scans</td>
</tr>
<tr>
<td>Bejan et al. [7]</td>
<td>MB-based, UMLS-concepts as features</td>
<td>Identifying IC patients with pneumonia from narratives</td>
</tr>
<tr>
<td>Ehrentraut et al. [16]</td>
<td>ML-based, using feature-selection</td>
<td>Detecting HAIs from Swedish patient records</td>
</tr>
</tbody>
</table>

### 3. Methods I: data collection and annotation

This section describes the corpus of reports and annotation process used for training (and evaluating) the various classifiers for identifying scan-reports and patients considered to be positive for IFD. Supervised machine learning techniques were used for the report classification task. A gold standard dataset of reports was constructed from a collection of CT reports sourced from three different hospitals (listed below) and annotated by experts in Invasive Fungal Disease.

#### 3.1. Data collection

CT reports were collected for patients known to have contracted an Invasive Fungal Disease (IFD) along with those for control patients (i.e., patients not diagnosed for IFD), from three different hospitals in Melbourne (Australia): the Alfred Hospital (AH), the Royal Melbourne Hospital (RMH), and the Peter MacCallum Cancer Centre (PM). IFD and uninfected patients from 2003 to 2011 inclusive were identified from previously completed observational and controlled clinical mycology studies [12,4,2,3,33]. Additional cases were identified by the physicians by interrogating HSCT (RMH), infectious diseases consultation (PM) databases and microbiology records (AH, RMH). Patients lacking CT reports were excluded. The formal inclusion and exclusion criteria are specified in a companion paper: [5].

CT reports were obtained from each hospital information system as text files and de-identified—i.e., manually stripped of any potentially identifying information such as name, address, age/birthdate, gender. CT reports of any anatomical site were included; however, in order to focus on sino-pulmonary involvement, brain scan reports were later excluded from analysis due to few patients with brain infection, except for brain scan reports performed in combination with another site e.g. chest and/or sinus. All scans were performed during the clinical encounter; an encounter was defined as being from admission to either discharge, death or transfer; or for those few outpatients, from performance of the diagnostic scan and for 12 weeks thereafter were included. A follow-up duration of 12 weeks was chosen since clinical practice guidelines recommend 12 weeks of treatment for IFDs [49]; evaluating radiological response over this period was therefore important.

For each report, we used the free-text section, which contains the radiologist’s interpretation of the scan and (with the exception of reports from one hospital) the reason for the requested scan as...
written by clinicians. We obtained data for 553 patients of the hospitals over the given period, i.e., over the total pool including IFD-positive and control patients. Each report was annotated independently of whether the patient had an IFD or not; therefore, in some cases a report from an IFD-positive may not be suggestive of IFD and would therefore be annotated as negative for such. Our goal for this work was to build and evaluate classifiers according to physicians’ interpretation using only the text in the scan reports, regardless of how other evidence (e.g., from pathology reports) may have impacted diagnosis of patients.

### 3.2. Data annotation

Manual annotation of the text reports was guided by annotation guidelines developed at the start of the process and subsequently refined through consultation. From the dataset of 553 patients, an initial sample of 73 IFD-positive patients was extracted in order to develop the annotation schema and to measure inter-annotator agreement.

The full set of annotation categories agreed on by the clinician experts is shown in Table 3 (see Appendix A); categories include both positive and negative presence of evidence of fungal infection, as well as indicators that may be appropriate to related tasks. Annotation at the concept-mention level was avoided in order to keep the annotation process relatively lightweight so as to make the process more easily transferable to other similar tasks. Annotation was performed by the three infectious diseases physicians (authors MAR, MS, KT) at both sentence and report level. Note that each sentence could have multiple labels, but reports had only one. Examples of annotated sentences are provided in Table 4 (Appendix A). The complete dataset was annotated by the primary annotator (MAR), with the other two annotators (MS, KT) annotating half the collection each, ensuring that all items were annotated by two experts. Inter-annotator agreement was measured using Cohen’s Kappa metric [8], the accepted standard for measuring agreement in linguistic annotation. Kappa scores were calculated by comparing the results of the primary annotator to those of the other annotators.

Sentence-level annotations were used to train and evaluate sentence classification systems, while report-level annotations were applied for evaluating the main task of report-level classification. We used the following subset of categories for creating classifiers; these were the categories that displayed the most reliable inter-annotation agreement, as indicated by their Kappa scores [28].

- **sentence-level category 1**: sentence that positively and specifically indicates likely IFD (Kappa = 0.64);
- **sentence-level category 3**: sentence that is a negative indicator of IFD (i.e., suggests no IFD) (Kappa = 0.58);
- **report-level category 3**: report that is overall negative for IFD (Kappa = 0.83).

We thus developed and evaluated binary sentence classifiers to distinguish two categories: “positive and specific to IFD” and “important negative for IFD”. Sentences were thereby classified into three mutually-exclusive categories: positive, important negative, or neutral with respect to IFD. At report level, we developed binary classifiers to discriminate negative reports from those that were positive or equivocal for IFD.

### 3.3. Gold standard and evaluation datasets

For classifier training and evaluation purposes, we partitioned our entire data collection into the following subsets, variously used for training sentence-level classifiers, and evaluating classifiers for the sentence-, report-, and patient-level classification task. Table 2 summarises these partitions and their purposes.

#### 3.3.1. Development Dataset

After finalising the annotation schema, we proceeded to annotate our gold standard data. From the initial collection of 73 patients, 13 patients were removed as they did not satisfy the inclusion criteria. Specifically excluded were patients that had brain scans only, not performed in combination with a lung or sinus scan. (Note: these patients were moved to the set of unannotated patients used as held-out testing data, as described below.) For cases where there was inter-annotator disagreement, annotations from the primary annotator (MAR) were used as gold standard, and the annotated dataset was subsequently extended by employing this annotator only. Similarly, reports from 60 control patients were annotated as per the process described above, resulting in a dataset of 120 patients annotated at both sentence- and report-levels. We henceforth refer to this as the Development Dataset: this corpus was used for developing the sentence-level classification models and for cross-validation evaluations. During the annotation process, we observed that some (8 in total) of the patients in the control set had reports that were annotated as IFD-positive by the medical expert; these patients were moved to the IFD-positive group. Conversely, there were two cases where IFD-positive patients did not have any positive evidence in the text, and these were moved to the control set. There was a total of 13 reports associated with these 10 patients. Table 6 shows characteristics of the Development Dataset: we see that control patients tend to receive fewer scans, and that there are clear differences across hospitals in the ratio of scans per patient.

#### 3.3.2. Held-Out Verified Dataset

An additional 40 patients (20 controls and 20 positives) were annotated at report level (not at sentence level), in order to create a held-out dataset for more reliable evaluation of our systems at report and patient level. It was found that 5 patients originally categorised as IFD-positive actually had no IFD-positive reports, and 2 control patients actually had reports labelled as IFD-positive by the medical experts; these patients (with 22 associated reports) were redistributed accordingly. Characteristics for this test set are also shown in Table 6. We henceforth refer to this dataset as Held-Out-Verified.

#### 3.3.3. Unverified data

The remaining reports were not manually labelled by the medical experts, neither at sentence nor report level: these remained as attributed to positive or control patients as determined by the treating hospital—i.e., diagnosed with IFD or not, determined as described in Section 3.1. These unlabelled reports were used for a final patient-level evaluation of our systems. The distribution across hospitals of this dataset is also shown in Table 6. In this case, it is possible that information used to classify patients as positive came from a source other than the reports themselves (e.g., pathology results), making it a less reliable gold-standard for purely text classification evaluation. This corpus is henceforth referred to as Held-Out-Unverified.

### 4. Methods II: constructing the classifiers

Following is a description of construction of the Text Classifiers, at each of sentence-, report-, and patient-levels.

---

1 These patients may have been diagnosed as IFD-positive using other tests; however, the scope of this paper is limited to the evidence in the scan reports themselves.
Table 2
Subdivisions of the dataset and their purpose.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Contents</th>
<th>Purpose</th>
</tr>
</thead>
</table>
| Development Dataset (annotated at sentence level) | Reports from 120 patients (60 IFD-positive, 60 control) | - Training sentence classifiers
- Evaluation at all levels (using cross-fold validation) |
| Held-Out-Verified Dataset (annotated for IFD at report-level) | Reports from 40 patients (20 IFD-positive, 20 control) | - Training report-level classifiers
- Report- and patient-level eval. |
| Held-Out-Unverified Dataset (held out; unexamined) | Reports from 393 patients | - Patient-level evaluation |
| Complete Dataset | 553 patients 1716 scan reports | |

4.1. Feature representations

For the implementation of our machine-learning approach to text classification, we identified a number of useful feature types, all of which were computed fully automatically into feature instances. Two main types of features were explored: “Bag” and “Structural”. Bag features could be applied to both sentence and report-level prediction, while Structural features could only be applied at the report level.

In order to represent the instances of the “Bag” features as feature vectors, lexical and semantic information within sentence boundaries was extracted. Documents were segmented into sentences using the JulieLab automatic sentence segmentor [44]; semantic information was obtained by applying the MetaMap parser [6], which maps phrases in text into medical concepts from the Metathesaurus of the Unified Medical Language System (UMLS) from the US National Library of Medicine (NLM).

Three subtypes of Bag features were tested: BOW is the basic bag-of-words model; the other two subtypes were constructed by adding semantic features:

- **Bag-of-Words (BOW):** The Genia Tagger [47] was used to tokenize text and remove punctuation. Dates were normalised into a “DATE” feature and numbers into a “NUMBER” feature. Finally, the position of question marks—beginning, inside, or end of the sentence—were marked to help identify speculative sentences.

- **Bag-of-Phrases (BOP):** Phrases mapped into UMLS medical concepts by MetaMap were collected for each sentence. Negated concepts were identified using Negex [10], and automatic disambiguation of concepts was performed using the module within MetaMap designed for this purpose. The identified phrases were marked with whether the concepts were found in positive or negative context.

- **Bag-of-Concepts (BOC):** Use of MetaMap, which links identified phrases to concept identifiers from the UMLS Metathesaurus, enabled generalisation by linking phrases with shared meaning. e.g., the terms “mycosis” and “fungal infection” are both linked to the concept “C0026946”. Bag-of-concept IDs detected in each sentence were identified, and again marked with whether the concepts were found in positive or negative context.

Structural features operated on the output of sentence classification and were used for classification of scan reports. There were two main types of Structural features, which were integrated into a single feature vector representation:

- **Scan type:** the anatomic site(s) of the target scan (e.g., “Chest scan”) were identified using regular expressions over the text of the reports. An additional feature was introduced for each scan-type to measure how many scans of the given type preceded the current one (e.g., “Number of previous chest-scan reports”).

- **Sentence-level classifications** (over previous and current reports): these features indicated the number of sentences that were scored with positive/negative labels in the current report (e.g., “Positives = 2”), and accumulatively over previous reports for that patient (e.g., “Accumulated Negatives = 4”). This set of features included a feature type indicating the latest positive/negative prediction made by the sentence classifier over the current document (e.g., “Latest-sentence-prediction = negative”).

4.2. Classification techniques

Different types of systems were developed to classify sentences and scan reports, mainly using Machine Learning (ML) techniques. Also evaluated were the use of rules (both hand-crafted and automatically derived), hybrid systems (combining ML and rules), and the use of simple heuristics.

4.2.1. Sentence-level classification

For sentence-level classification, the task was to perform 3-way classification, i.e., to determine whether each sentence in a report provided positive, negative, or neutral evidence for IFD. This task was modelled as two separate binary classification problems: i.e., discrimination between IFD-positive sentences and others, and between IFD-negative sentences and others. Keyword-matching classifiers and ML approaches were used at this level; these approaches were evaluated separately.

**Keyword-Matching Approach:** As a baseline approach, we devised a simple method to classify sentences as positive for IFD using a set of keywords compiled as described below. Another list of terms was used to identify when a given sentence contained negation; in that case, or when the keyword was followed by a question mark, the sentence was marked as neutral.

We first created a list of positive-indicator terms compiled by medical experts from the literature [21]; these terms were mapped to UMLS concepts. The final list of manually curated terms is shown in Table 7.

As a variation to the above, we explored automatic identification of a list of terms to be used for keyword-matching, leveraging the feature set defined in Section 4.1; in particular, features were scored according to their log-likelihood of appearance in the collection. For a given term, log-likelihood was calculated as per Eq. (1), with additive smoothing applied and $\epsilon = 0.1$ [1]:

$$\text{Loglikelihood(feature)} = \log \frac{\text{Probability(IFD|feature)}}{\text{Probability(NEG|feature)}} + \epsilon$$  \hspace{1cm} (1)

Terms scoring above a given threshold were selected as indicative of IFD, and cross-validation was used to avoid overfitting the data. We present results for different thresholds in Table 8, which shows the top-scoring terms with log-likelihood value (over the full dataset) greater than 2.6.

**Machine Learning Approach:** The second set of classifiers used standard ML techniques. A number of characteristics of the training data led us to experiment with a variety of ML algorithms. Due to “Bag” features being used for sentences, these techniques were required to cope with large and sparse feature vectors. The fact that most sentences in the training data were neutral—i.e., neither negative or positive—also contributed to the sparseness. Moreover, the cost of failing to identify positive sentences is higher
than missing neutral or negative sentences, leading us to bias towards high Recall over Precision.

Machine learning algorithms used include: Support Vector Machines (SVM); Bayesian Network models; Naive Bayes; and Random Forests. For each of these algorithms, we also applied feature selection to handle the large amount of feature instances. Specifically, we used a correlation-based feature subset selection method, which considers the individual predictive ability of each feature and the redundancy of each subset [23]. Our configuration used Best-First search, with a cache-size of one element, and five levels of backtracking.

The Weka Machine Learning toolkit [50] was used for the implementation of all the ML algorithms; the referenced book contains extensive documentation about the techniques used.

**Hybrid Approach:** For the task of identifying IFD-positive sentences, we experimented with combining the outputs of the ML classifier with the automatic keyword extraction technique, using a **Conservative** rule: if either of the methods assigned a positive label to the target sentence, it was labeled as IFD-positive.

### 4.2.2. Report-level classifiers

For report-level IFD classification, heuristic rules and machine learning approaches were used to perform binary classification (positive/speculative vs. negative).

**Heuristic Rules Approach:** This approach leveraged the sentence-level classifiers described in Section 4.2.1 by applying simple rules over the sentence-level classifiers in order to minimise false negatives and achieve high Recall. Two candidate heuristic rules were defined as follows:

- **Conservative:** label a report as IFD-positive if any sentence in it is labeled IFD-positive;
- **Balanced:** label a report as IFD-positive when it contains more sentences labelled as IFD-positive than IFD-negative.

**Machine Learning Approach:** ML was applied directly to report classification, using the methods described in Section 4.2.1. For this task, the two classes of features presented previously—i.e., “Bag” and “Structural”—were used; note that one of the “Structural” feature types was the output of sentence-level classification.

In order to maximise Recall, we also combined the outputs of the best ML approaches voting in a conservative way, i.e., assigning a label indicating IFD-positive whenever a sentence-level classifier did so. Both “Bag” and “Structural” feature-based methods were tested. Also evaluated was performance using the top-k systems (both in precision and recall) using “Bag” features, for various k.

Note that we only report here results for report-level classification that used the best-performing sentence-level techniques.

### 4.2.3. Patient-level classification

A simple heuristic was used for patient-level classification: a positive label was assigned to a patient if any of their reports was classified as positive. This was done so as to minimise false negatives at this level, i.e., missed patients with IFD. Similarly to report-level classification, we report only those patient-level results that use the best-performing sentence- and report-level techniques.

### 4.3. Evaluation framework

We used the standard accepted metrics of Precision, Recall, and F-score to evaluate the performance of the various binary classifiers constructed. High Recall is crucial at report and patient levels; it is also likely to be an important factor when identifying sentences that are positive for IFD. High Precision is also desirable, to avoid the cost of false positives.

For our evaluation of sentence classification, 10-fold cross-validation over the Development Dataset was applied. We evaluated classifier performance at sentence, report, and patient levels. We then trained our classifiers over the Development Data (classified at sentence-level), and tested over the Held-Out-Verified (annotated at report level), measuring performance at report and patient levels. Finally, systems were trained over the Development Dataset and evaluated at patient level over the Held-Out-Unverified, which could only be used to evaluate at patient level.²

### 5. Results I: classification over the Development Dataset

Presented here are results for the classification tasks over sentences, scan-reports, and patients, using the gold-standard Development Dataset of 120 patients, using 10-fold cross-validation.

#### 5.1. Sentence-level classification

As discussed in Section 3, binary classification was performed at sentence level for the labels “positive and specific to IFD” and “important negative for IFD”.

**Baseline.** Table 9 shows performance results for baseline systems that used a keyword-matching approach, as described in Section 4.2.1. Rules using the manually-defined list of IFD-indicative keywords resulted in high precision (0.850) but low recall (0.434). Rules using the automatic log-likelihood technique to construct lists of IFD-indicative keywords achieved maximal F-score (0.639) when using a log-likelihood threshold of 1.6 for term-extraction.

**ML Classifiers for Positive Indicators.** Having tested the keyword-based methods, we explored the use of different Machine Learning (ML) classifiers and feature sets over the Development Dataset.

The first columns of Table 10 show results for all systems. In the top row, we include a naive “Majority Class” baseline, which consists of always assigning the class most frequently seen in the training data.

Feature Selection (see Section 4.2.1) was required when using Bayesian Networks and Random Forest due to their computational costs. The scores illustrate that Naive Bayes provided high Recall (0.858 using BOC features), but at the cost of very low Precision (0.338 at this Recall). The next-best classifier with regards to recall was SVM (0.690 using BOC features), with considerably better Precision (0.715 at this recall) than for Naive Bayes. SVM produced the highest overall F-scores (0.705) using BOC features. Feature Selection did not improve performance when combined with SVMs. Given these results, we chose SVM as the positive sentence-level classifier when performing classification at report and patient levels.

**Classifying for Negative Indicators.** We evaluated the performance of binary sentence classification for the label “important negative for IFD”, using only the ML approach.

The central columns of Table 10 show the performance results for the different configurations. For this task, lower Recall may result in more false positives at the report level, which is regarded not as harmful as false negatives for our task setting. We therefore prefer Precision over Recall for this class. SVM classification using BOP as features was the best performing in terms of both Precision (0.773) and overall F-score (0.772). Feature Selection was again detrimental for SVMs. This further supports the choice of SVMs for sentence-classification for use in the report- and patient-classification tasks.

² Recall that patients were pre-known to have been diagnosed with IFD or not.
5.2. Report-level classification

Report-level classification, i.e., classifying scan reports as IFD-positive or IFD-negative, is a critical task for surveillance, with Recall as the most important metric: a false negative at the report level may translate into a missed opportunity to raise an early alarm of the presence of an IFD.

Direct Document Classification. The first experiment evaluated a traditional document classification approach: report-level features were extracted and used to build models and generate predictions. As done with sentence classification, we present a “majority class” baseline together with the classifiers, in order to know the expected precision from the simplest of methods.

Results for various classifiers are shown in the rightmost columns of Table 10. The highest Recall obtained was 0.847, using SVM and BOP features: i.e., almost 15% of positive reports would be missed. Precision and Recall were balanced for the various systems, whereas the aim was to maximise Recall. The naive baseline approach—which labelled every report as positive—itself produced an F-score of 0.741, which was similar in performance to the weakest ML configurations.

Using Sentence-Classifier Output. We sought to improve report-level performance by leveraging the output of the sentence-level classifiers and applying the heuristic rules (Conservative and Balanced) defined in Section 4.2.2.

For sentence classification, the best-performing ML approaches for positive- and negative-sentence classification were used: these were SVMs with different feature sets. Table 11 presents the results of using SVMs with BOP as features. When using the Conservative heuristic we obtained 0.893 recall and 0.810 precision.

Alternative hybrid approaches were also applied: these combined learned rules with the use of ML systems. As per Section 5.1, using different log-likelihood thresholds for extracting relevant terms resulted in different term-matching rules. Table 11 shows results obtained when using log-likelihood thresholds in the range 1.4–1.8, which achieved the best performance for sentence classification. Recall of 0.930 and Precision 0.790 were obtained when using 1.8 as log-likelihood threshold.

Overall, use of rules over sentence-classification output outperformed the direct document-classification approach, and the results suggest that the Conservative heuristic is preferable to the Balanced one, particularly for maximising Recall.

The use of Structural features was also explored—i.e., use of scan-type as a feature combined with the output of sentence classifiers used as features rather than as the basis of the heuristic rules. SVMs (using BOP features) were used to classify sentences. Different classifiers were then applied at the report level, using sentence-classification results as features, along with the other structural features described in Section 4.1. The results are shown in Table 12. Bayesian Networks performed best, with F-score of 0.844, but with lower recall (0.842) than our previous approaches.

Ensemble Classifiers. Finally, ensemble classifiers were evaluated by combining the outputs of the best ML approaches voting in a Conservative way, i.e., by assigning a positive label to a report whenever any classifier at the sentence-level assigned a positive label. Both Bag and Structural feature types were tested. Performance when using only the top-k systems (for various k) for Bag features was also tested. The results are shown in Table 13. The use of top-3 systems reached 0.930 Recall and 0.779 Precision. Recall improved to 0.995 when systems using Structural features were added, although with a drop in Precision (0.700). Table 14 summarises the best approaches for report-level classification over the Development Dataset. The results indicate very high Recall using various methods (0.995 when using all features). The highest F-score (0.855) was achieved using the Conservative heuristic for report-classification over SVM sentence predictions, with 1.8 as the log-likelihood threshold for term-extraction for creating keyword-matching rules.

5.3. Patient-level classification

The final set of evaluations over the Development Dataset was performed for patient-level classification, to test the ability of the classifiers to discriminate IFD-positive patients from control patients. Achieving high Recall was the overriding aim of this task. Thus, a Conservative approach over our different report-level classifiers was used: if any scan report for a given patient was classified IFD-positive, then the corresponding patient was also labeled IFD-positive. Table 14 shows patient-level results using the main methods evaluated for report classification.

Two of the systems achieved near-perfect Recall of 0.985 ("ML Combine All" and "Report Rules – Conservative Sentence SVM, thr = 1.8"): each of these systems produced only a single false negative. The false-negative patient was different in each case, but each patient had few (one and three) scan reports, each of which was short and came from the same hospital (Royal Melbourne Hospital); this made it more difficult for classifiers to find positive indicators. Highest Precision was achieved for Bayesian Networks with Structural features (0.734).

6. Results II: evaluation over the Held-Out Datasets

Presented here are results from testing over the two Held-Out datasets described in Section 3, i.e., Held-Out-Verified (containing 40 patients annotated at report-level) and Held-Out-Unverified (containing 393 patients, unannotated); these datasets were not used during training. System configurations evaluated at report and patient levels using the Held-Out-Verified data, and at patient level only for Held-Out-Unverified.

Report-Level Classification. Table 15 shows a summary of report-level classification performance over the Held-Out-Verified test set. As with the Development data (refer to Table 14), the use of Conservative heuristics with a threshold of 1.8 was again best performing, with improved Recall of 0.997, at a cost to Precision (best of 0.731). The overall F-scores were slightly lower (highest of 0.812), but the improvement over the baseline was greater, due to the better balanced distribution of IFD-positive and control reports in this dataset. Overall, we demonstrated high recall (0.950) at acceptable precision (0.70).

Patient-Level Classification. Table 15 shows performance results at patient level over the Held-Out-Verified test dataset. In this case, most system configurations obtained perfect recall (1.0), which translates into no IFD-positive patient being missed. Precision was slightly lower compared to evaluation over the Development set (best of 0.654 vs. 0.734) (refer to Table 14).

Finally, the systems were evaluated over the Held-Out-Unverified test dataset, for which only patient-level classifications are known: i.e., this dataset contains reports that were not manually classified by our medical expert annotators, and all that is known is the original classification of the patients as either known IFD-positive or control. Table 16 shows the results of performing patient-level classification over this dataset. The systems once again displayed very high Recall. Best performance trade-off between Recall and Precision was achieved by the ML system classifying reports directly using BOP features, without relying on sentence classifiers (F-score: 0.799); next-best were systems using Conservative heuristics over sentence-classification (F-score: 0.756).
Overall, the configuration using Conservative heuristics over sentence classifiers was the most robust over the most reliable datasets (Development and Held-Out-Verified). Performance of this system over Held-Out-Unverified patient-level data was examined in more detail. This configuration, combined with a log-likelihood threshold of 1.8, achieved a Recall of 0.979, equivalent to five false negatives. Manual review of the reports for these missed cases found that none of them contained any positive indications for IFD in the text—i.e., these were patients that were most likely assigned a diagnosis of IFD using clinical evidence beyond the CT scans.

7. Discussion

We have explored various architectures for classifying sentences, scan reports, and patients into useful categories for IFD surveillance tasks that leverage information in CT scan reports. Over different datasets, our techniques produced very high Recall at both report and patient level; while the Majority Class baselines also reached perfect Recall (not uncommon for such a baseline), our techniques produced Precision scores well above those baselines. In particular, the intermediate step of classifying sentences into categories proved beneficial to report classification, an extension to standard approaches to this task (c.f., [7,16]). Of note is that even with Recall of less than 0.7 for SVMs on the (positive) sentence-classification test (Table 10), this still contributed to very strong Recall for report- (0.930) and patient-level (0.985) classification (see Table 14). In a practical Decision-Support System, sentence classification could also be useful for highlighting pertinent evidence for the benefit of the clinical decision-making user.

Regarding the best performing methods, at report level the strongest performance was achieved using Conservative rules over sentence-classification output. The use of log-likelihood calculations for automatically deriving keywords performed robustly over the Development and Held-Out-Verified datasets, resulting in Recall of 93% and 95%, and Precision of 79% and 71%, respectively.

Results for patient-classification confirmed the superior performance of our “hierarchical” classification approach (whereby Conservative report-level classification rules make use of ML sentence-level classification), where Recall scores of 98% and 100% were achieved over Development and Held-Out-Verified datasets respectively (corresponding Precisions were 69% and 63%). Interestingly, the use of log-likelihood thresholding benefits the performance at report level, but not at patient level, where the Recall was already high. The final results over the patient-level annotated data (Held-Out-Unverified) were similar, but our manual analysis of false positives showed that this data is not completely reliable as a gold-standard for measuring performance of patient-level classification: some patients seem to have been diagnosed with IFD using clinical evidence beyond the CT scan reports.

These overall results, and error analysis of false negatives, indicate that we have largely achieved our goal of avoiding missed cases of IFD. Over the Development set, only a single positive patient was missed by the best-performing report-level system; it is important to note that this patient had only a single scan, and agreement between human annotators was not perfect at report level (see Section 3). For the Held-Out-Verified set, no IFD-positive patient was missed by the classifier.

Precision, which quantifies the number of “false alarms”, was acceptable but substantially lower than Recall. Although high Recall is the more important characteristic to the surveillance task, minimising false positives is nevertheless desirable so as to avoid “alert fatigue”. The drop in Precision is particularly evident in experiments over the Held-Out datasets. Analysis of the false positives suggests that this may be due to subtleties in the annotation: more highly speculative sentences in these examples were annotated as negative by the physicians, leading to a lack of penalty to Recall when these were missed but a penalty to Precision if they were classified as positive. One approach to addressing the problem would be to develop a more sophisticated treatment of speculative language [48].

Regarding the impact of Precision at report level, we note that not all false positives have the same impact in our setting. If a patient has already had a scan flagged as positive by the system, in a practical setting an action to monitor the patient should follow. Subsequent scans marked as positive for the same patient should imply less overhead in a monitoring system than the alerts for patients without prior positive scans. Another interesting situation arises when a false positive refers to a patient that will eventually have a positive scan (even if the current scan is negative). In this case the classifier may have found an early indicator of risk, and the effort of monitoring this patient may be worthwhile.

We examined the distribution of the positive scans from our most robust classifier in order to better understand the impact of the different types of positive scans. Our evaluation showed that our best system performs at 79% Precision at report level over the Development set; i.e., 53 of the 253 (21%) predicted positive scans are false positives. In 26 cases (49% of all false positives), the system indicated a prior scan for that patient as positive. In 13 cases (25% of all false positives), the false positive was an early alert for a patient. We manually examined the early indicators and found a large number of sentences that our annotators had considered “Positive but not specific to IFD” (see Section 3); e.g., “A focus of fungal infection cannot be excluded”. Thus, for only 14 cases (26% of the false positives) was there an alert for a patient that ends up being negative for IFD.

Over the Held-Out-Verified set (Precision 71%), we received 17 false positives from the 58 scans that were predicted as positive (29%). The number of cases where the system predicted a previous positive scan for the given patient was 9 (53% of all false positives), and the number of early indicators in this case was 0; therefore only for 8 cases (47% of the false positives) was there an alert raised for a patient that did not end up being IFD-positive. As indicated earlier, such cases were often caused by a sentence containing “speculative” language, e.g., “Query IFD”: the clinician annotators would mark the sentence as “neutral” but the system classified the sentence as “indicative of IFD”.

8. Conclusions

This paper has explored the task of analysing the text in radiology reports, specifically of CT-scans, for evidence suggestive of an Invasive Fungal Disease (IFD). We developed and evaluated a suite of innovative techniques, combining sentence- and report-level ML-based classification combined with automated keyword extraction using log-likelihood measures. In particular, our techniques required a lightweight annotation, requiring annotation only at the sentence level. The resulting text classification systems—developed from reports derived from three major Melbourne hospitals—achieved high Recall (i.e., minimal missed incidence) at both report and patient levels. High Recall is an important attribute of a pervasive surveillance system intended to screen patients in order to identify those with potential fungal infection. Invasive fungal diseases are challenging to definitively diagnose, due to poorly sensitive tests compounded by patients often being too unwell to undergo invasive diagnostic procedures. Automatically mining the narrative of CT reports offers the
possibility of eventually making real-time surveillance of these infections feasible for hospitals.

The task of surveillance using CT reports requires accurate discrimination of patients with IFD from those without IFD, and in particular, minimal missed cases. The Recall achieved with the best-performing classifiers at report-level and especially at patient-level was very high, with all patients in our test data with an abnormal scan (as determined by the clinical experts) being successfully identified at some point during their clinical encounter.

The report classification task underpinned the patient-level surveillance but is also a potentially useful task in its own right. Real-time prospective continuous surveillance via report processing could be used to: link to clinical decision support through the triggering of treatment protocols; alert hospitals to potential outbreaks; strengthen antifungal stewardship in hospitals; assist clinicians in evaluating the effectiveness of preventative strategies such as antifungal drug prophylaxis or infection prevention practices; inform clinical trial design with contemporary incidence rates; and provide an opportunity for tracking disease trends in response to rapid advances in medical and transplant practices.

Future work will concentrate on prospective evaluation of the classifier in clinical environments with broader engagement from end-users. We will also continue improving our classification techniques, including performing statistical analysis of actual distributions of positive examples as they arise in practice, and evaluate classifier performance under simulated hospital conditions.

Acknowledgments

We thank the anonymous referees for several extremely helpful suggestions for improving the presentation of this work.

This work was performed while Martinez and Cavedon were employed at NICTA Victoria Research Laboratory. NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

Appendix A Supplementary material

Supplementary data associated with this article can be found in the online version, at http://dx.doi.org/10.1016/j.jbi.2014.11.009.

References


