original report

Toward Electronic Surveillance of Invasive Mold Diseases in Hematology-Oncology Patients: An Expert System Combining Natural Language Processing of Chest Computed Tomography Reports, Microbiology, and Antifungal Drug Data

Michelle R. Ananda-Rajah
Christoph Bergmeir
François Petitjean
Monica A. Slavin
Karin A. Thursky
Geoffrey I. Webb

Michelle R. Ananda-Rajah, Alfred Health; **Michelle R. Ananda-Rajah**, **Christoph Bergmeir**, **François Petitjean**, and **Geoffrey I. Webb**, Monash University; and **Monica A. Slavin** and **Karin A. Thursky**, Peter Doherty Centre for Infection and Immunity; University of Melbourne, Melbourne, Victoria, Australia.

Supported in part by the Monash Institute of Medical Engineering seed fund 2015.

Corresponding author: Michelle R. Ananda-Rajah, FRACP, PhD, General Medical Unit & Infectious Diseases Unit, Level 6 Alfred Centre, Alfred Health, 85 Commercial Rd, Melbourne, Victoria 3004, Australia; e-mail: michelle.ananda-rajah@monash.edu.

abstract

Purpose **Prospective epidemiologic surveillance of invasive mold disease (IMD) in hematology patients is hampered by the absence of a reliable laboratory prompt. This study develops an expert system for electronic surveillance of IMD that combines probabilities using natural language processing (NLP) of computed tomography (CT) reports with microbiology and antifungal drug data to improve prediction of IMD.**

Methods **Microbiology indicators and antifungal drug–dispensing data were extracted from hospital information systems at three tertiary hospitals for 123 hematology-oncology patients. Of this group, 64 case patients had 26 probable/proven IMD according to international definitions, and 59 patients were uninfected controls. Derived probabilities from NLP combined with medical expertise identified patients at high likelihood of IMD, with remaining patients processed by a machine-learning classifier trained on all available features.**

Results **Compared with the baseline text classifier, the expert system that incorporated the best performing algorithm (naïve Bayes) improved specificity from 50.8% (95% CI, 37.5% to 64.1%) to 74.6% (95% CI, 61.6% to 85.0%), reducing false positives by 48% from 29 to 15; improved sensitivity slightly from 96.9% (95% CI, 89.2% to 99.6%) to 98.4% (95% CI, 91.6% to 100%); and improved receiver operating characteristic area from 73.9% (95% CI, 67.1% to 80.6%) to 92.8% (95% CI, 88% to 97.5%).**

Conclusion **An expert system that uses multiple sources of data (CT reports, microbiology, antifungal drug dispensing) is a promising approach to continuous prospective surveillance of IMD in the hospital, and demonstrates reduced false notifications (positives) compared with NLP of CT reports alone. Our expert system could provide decision support for IMD surveillance, which is critical to antifungal stewardship and improving supportive care in cancer.**

Clin Cancer Inform. **© 2017 by American Society of Clinical Oncology**

INTRODUCTION

Invasive mold disease (IMD) has significant health and economic consequences[1-3] and is likely to increase as the population of immunocompromised patients expands.[4] IMD most often presents as pneumonia in patients with impaired immunity due to a variety of causes, including chemotherapy for cancer, immunomodulatory drugs, or transplantation.[3,5,6] IMD is associated with mortality rates of > 50% in patients with hematologic malignancies and transplantation recipients[3,6] and may adversely affect long-term cancer cure rates due to modifications in curative chemotherapy regimens.[7] Hospitals spend millions on antifungal drugs but have not invested in prospectively monitoring their local epidemiology (ie, institution-specific pathogens, rates, and patients affected), despite it being a key performance indicator of antifungal drug stewardship programs.[8] Knowledge of local fungal epidemiology is important for several reasons. It informs local guidelines,[8,9] infection prevention efforts, clinical trial design,

clinical audit, and benchmarking that addresses interfacility clinical variation.[10] Epidemiologic surveillance of IMD is difficult because of the absence of a consistent laboratory prompt, such as a positive blood culture. As a result, case finding relies on manual interrogation of multiple data sources, including bedside clinical review, radiology, and microbiology, which makes it a costly, onerous task not performed by hospitals outside research protocols.[3,5,6] As a result of an absence of surveillance data, hospitals are ill equipped to evaluate their clinical practice, rationalize antifungal drug use, detect outbreaks, or identify new, previously under-recognized at-risk patient groups.

To facilitate monitoring of IMD in hospitals, we developed natural language processing (NLP) of chest computed tomography (CT) reports.[11,12] We have previously shown that the primary screening method for IMD, with a high sensitivity to maximize case finding, should focus on chest CT.[11,12] CT is universally performed when IMD is suspected, and pulmonary involvement is present in 90% to 100% of patients with IMD.[3,5,13] Although lung sampling (ie, biopsy or bronchoalveolar lavage [BAL]) and other laboratory indicators (eg, sputum culture, biomarkers) could be used for epidemiologic surveillance, they are not performed with the same frequency as CT scans.

Our baseline text classifier, detailed elsewhere,[11,12] was based on automatically learned, hierarchically organized, cascading text categorization techniques. In brief, we obtained 1,880 free-text CT reports for 270 patients with IMD and 257 control patients from three tertiary hospitals.[11] We analyzed IMD evidence at patient, report, and sentence levels. Training data were obtained by three infectious disease physicians who annotated a subgroup of 449 reports from 79 patients with IMD and 68 control patients for language features suggestive of IMD.[11] We tested a variety of machine learning, 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.[11,12] The best system (using support vector machines) achieved high recall over unseen data, with a sensitivity of 91%, specificity of 79%, and receiver operating characteristic (ROC) area of 0.92, with few clinically significant missed notifications (0.9%) at report level,[11] whereas in a separate experiment that used a slightly smaller data set (n = 1,716 free-text CT reports), it identified 100% of patients with IMD.[12] With a negative predictive value of 97% to 99% (across a range of hypothetical IMD prevalence rates), IMD could be excluded with confidence.[11] Our next goal was to minimize false notifications (ie, improve specificity) while maintaining or improving case detection (ie, sensitivity). Therefore, we developed an expert system that incorporated derived probabilities from NLP of CT reports, medical knowledge rules, and machine learning classifiers for processing adjunctive data (microbiology and antifungal drug–dispensing data) to enable real-time, sustainable, and network-wide surveillance of IMD with minimal effort.

METHODS

Study Design and Setting

This was a retrospective case-control cohort study of patients from three tertiary adult university-affiliated hospitals (Alfred Health, Peter MacCallum Cancer Institute, and Royal Melbourne Hospital). Alfred Health and Royal Melbourne Hospital operate statewide allogeneic and autologous hematopoietic stem-cell transplantation (HSCT) services, which collectively perform approximately 250 allogeneic transplantations per year. Peter MacCallum Cancer Institute performs autologous HSCT only. The human research ethics committees at each site approved the study.

Clinical Data and Definitions

We selected a random convenience sample of 64 infected case patients and 59 uninfected control patients from the original data set used to develop the text classifier.[11] Microbiology and antifungal drug–dispensing data for each clinical encounter (defined from admission to discharge, death, or transfer from hospital) were extracted from hospital information systems and combined with probabilistic outputs of the text classifier derived from the earlier study.[11]

IMD was classified according to internationally accepted definitions.[14] Possible infections had suggestive radiologic features in the appropriate clinical context but lacked positive microbiology, whereas probable or proven infections had positive microbiology, such as isolation of a fungal pathogen from sputum or tissue.[14]

Microbiology included microscopy and culture of specimens from sterile and nonsterile body sites, including blood, BAL, tissue, pleural fluid, sputum, and biomarkers (galactomannan [GM], panfungal or *Aspergillus*-specific polymerase chain reaction [PCR]). We obtained pharmacy-dispensing data for the following antifungal drugs—fluconazole, voriconazole, posaconazole, liposomal amphotericin, caspofungin, itraconazole, and terbinafine—as well as the dosage, formulation (oral, intravenous), presentation (vials, tablets), and dates dispensed. None of the hospitals had electronic prescribing

systems in place, so drug-dispensing data approximated the actual dose administered to the patient.

Development of the Expert System

Our approach can be deconstructed into two phases: 1) We detected the patients at high likelihood of IMD by combining the probabilistic output of the text classifier with expert knowledge rules about IMD, and 2) for the remaining patients, we trained a classifier on all the available features including output of the text classifier, antifungal drug–dispensing data, and microbiology investigations and results.

Preprocessing

Preprocessing involved transforming the manually collected data from long to wide format. Each electronic event (ie, CT scan, drug dispensed, microbiology result) per clinical encounter was given a time stamp (date). Electronic events represented the following variables: probabilistic outputs of text classification of CT scan reports derived from the earlier study,[11] results of key microbiological investigations, and antifungal drug use (Table 1). Categorical variables (ie, positive, negative, or no test available) were assigned for each electronic event.

For example, electronic events were named Scan_Any positive, Blood_Any positive, BAL_Any positive, Pleural Fluid_Any positive, PCR_Any positive, GM_Any positive, Sterile_Any positive, and Non_sterile_Any positive, with "Any positive" denoting all positive results in that category for each clinical encounter. Scan_Any positive was set to positive if the text classifier returned a value $\geqslant 0.5$ with the highest probability from any scan in the clinical encounter supplied to the expert system. Isolation of *Candida* species from the pulmonary or oropharyngeal tract was not regarded as significant.

A clinical encounter with a negative CT scan by NLP and positive sterile site microbiology (ie, Sterile_Any positive) was marked as positive for IMD and not sent to the final classifier. Isolation of fungal pathogens from a nonsterile site, such as sputum, usually denotes colonization rather than invasive infection. However, for purposes of epidemiologic surveillance, we regarded the isolation of molds from sputum (ie, Non_Sterile_Any positive) as significant because of the high likelihood of either occult or subsequent invasive disease in this high-risk population.[15]

We created handcrafted rules on the basis of medical expertise (M.R.A.-R., K.T., M.S.) and informed by real-world data from a previous study[1] to interpret complex antifungal drug–dispensing data. These rules aimed to identify drug regimens more likely to reflect definitive treatment of IMD

Table 1. Definitions of Electronic Event Variables for CT, Microbiology, and Antifungal Drug–Dispensing Data Used by the Expert System

Electronic Event	Hospital Information System	Description
Scan_Any positive	Radiology	CT scan report with a probability > 0.5 for IMD by the baseline text classifier any time during the patient's clinical encounter (ie, episode of care).
Specimen_Any positive	Microbiology	Specimens can be from sterile or nonsterile anatomic sites. These may be samples from blood, tissue, pleural fluid, sputum, or wound sites. Any positive refers to isolation of a fungal pathogen at any time during the patient's clinical encounter (ie, episode of care).
Sterile_Any positive	Microbiology	Sterile sites include blood, pleural fluid, BAL, tissue (eg, lung or sinus biopsy). Any positive refers to isolation of a fungal pathogen at any time during the patient's clinical encounter (ie, episode of care).
Nonsterile_Any positive	Microbiology	Nonsterile sites include sputum, nasopharyngeal swab, mouthswill, or occasionally, wounds. Any positive refers to isolation of a fungal pathogen at any time during the patient's clinical encounter (ie, episode of care).
PCR or GM_Any positive	Microbiology	PCR and GM are molecular biomarkers for fungal pathogens. PCR may be performed on a variety of specimens, including blood and tissue. GM is usually performed on lung specimens (eg, BAL) or serum. Any positive refers to a positive result at any time during the patient's clinical encounter (ie, episode of care).
Rule 1	Pharmacy	Set to "yes" if voriconazole, irrespective of formulation, exceeded 4,200 mg per clinical encounter.
Rule 2	Pharmacy	Set to "yes" if at least 150 mg of liposomal amphotericin was dispensed daily for 7 days or more during a clinical encounter.
Rule 3	Pharmacy	Set to "yes" if 50 mg or more of caspofungin was dispensed daily for 5 days or more per clinical encounter

Abbreviations: BAL, bronchoalveolar lavage; CT, computed tomography; GM, galactomannan; IMD, invasive mold disease; PCR, polymerase chain reaction.

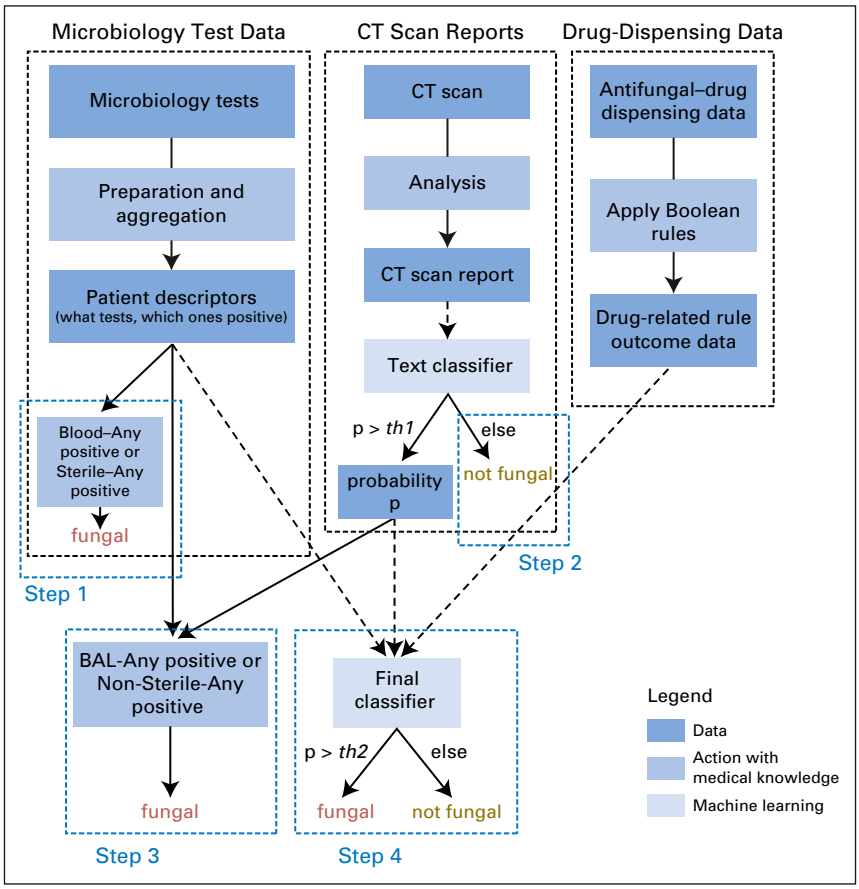

Fig 1. Architecture of the expert system incorporating natural language processing, microbiology, and antifungal drug–dispensing data for invasive mold disease classification. Sterile specimens include blood, pleural fluid, bronchoalveolar lavage (BAL), and tissue (eg, lung or sinus biopsy). Nonsterile specimens include sputum, nasopharyngeal swab, mouthswill, or occasionally, wounds. Any positive refers to isolation of a fungal pathogen at any time during the patient's clinical encounter (ie, episode of care). Th1 refers to the threshold of 0.5 that was prespecified for the baseline text classifier,[11,12] meaning that a report with a probability ⩾ 0.5 was flagged positive. Th2 is the threshold selected by the best-performing final classifier (Table 3) to maintain sensitivity and maximize specificity. Dashed lines represent inputs to the classifiers as opposed to processing steps and inputs to the expert system. CT, computed tomography.

rather than empirical or prophylaxis indications. Our earlier study[1] observed that median duration of inpatient therapy was 6 days for voriconazole, 7.5 days for liposomal amphotericin, and 6 days for caspofungin in patients with invasive fungal disease. Our drug rules were based on these observations but modified (Table 1), because drug dispensing from pharmacy to the hospital wards is not always consecutive, and the volumes of drug dispensed are variable. If drugs were given but not above the thresholds set by the rules, the variables were set to either "low" or "no." We did not include rules for fluconazole, itraconazole, and posaconazole, because these drugs are more frequently used for prophylaxis rather than treatment of established IMD, and our pharmacy dispensing data do not provide clinical indications for prescription. Terbinafine was not included in our rules because it is reserved for the rare episode of invasive scedosporiosis, where it is coadministered with voriconazole, a scenario that would have been detected by our voriconazole rule.

Architecture of the Expert System

The expert system (Fig 1) comprised machine-learning classifiers for text analysis and final classification, algorithms informed by medical knowledge, and clinical data. Fungal isolates from sterile sites, including blood (ie, Blood_Any positive or Sterile_Any positive), which unequivocally represented invasive disease were identified with a medical rule (Step 1, Fig 1). The baseline text classifier then screened all CT reports (Step 2, Fig 1). Positive IMD cases from the text classifier were classified as positive, without sending them to the final classifier, if either BAL_Any positive (ie, fungal isolation from lung) or Non_Sterile_Any positive (ie, fungal isolation from sputum) were labeled as positive. In this way, only undecided cases with negative or no tests on Blood_Any positive, BAL_Any positive, Sterile_Any positive, and Non_Sterile_Any positive were sent to the final classifier (Fig 1). Thus the final classifier processed probabilistic outputs from the text classifier of CT reports, drug data not subject to the aforementioned rules, and negative or missing microbiology.

Classification Algorithms

We experimented with several machine-learning algorithms, shown in Table 2. For implementation, we used the R programming language, together with Weka and the R Weka[16] package. All algorithms

Table 2. Parameters Used for the Machine Learning Algorithms

Algorithm	Parameters
Naïve Bayes	—
Logistic	MaxIts = −1, ridge = 1.0×10^{-8}
Random forest	numTrees = 10, numFeatures = 5
J48	confidenceFactor = 0.25, minNumObj = 2
SMO	c = 1, cachesize = 250,007, exponent = 1.0

Abbreviation: SMO, sequential minimal optimization.

used their Weka default configurations (the C4.5 implementation in Weka is called J48).

Evaluation Procedure

Evaluation was performed for the expert system overall. We used 10-fold cross-validation that used all available data to train and test the expert system, meaning that we obtained a prediction for each of the 123 data points. The results of the expert system for each clinical encounter (with IMD diagnosed or not) was compared with physician-adjudicated opinion on the basis of published diagnostic criteria,[14] which allowed for the calculation of sensitivity, specificity, and ROC.

RESULTS

Patient Characteristics

The data set of 123 patients comprised 64 (52%) patients with IMD and 59 (48%) uninfected control patients (Table 4). Neutropenia ($\leq 0.5 \times 109$ cells/L) was present in 86% and 71% of case and control patients, respectively, and was prolonged (median, 17 and 18 days, respectively). A history of HSCT was present in 53% and 46% of case and control patients, and was allogeneic in 85% and 78% of patients, respectively. IMD was probable or proven in 26 of 64 patients (41%). Sinus and/or pulmonary disease occurred in 90% of case patients. Invasive aspergillosis comprised 14 of 26 (54%) of microbiologically confirmed cases, with rare molds, including *Scedosporium* and *Rhizopus* species, identified in nine of 26 (35%).

Performance of the Final Classifier

Performance of the final classifiers for different threshold values is shown in Table 3 and Figure 2. Performance was best using naïve Bayes, which increased the specificity from 50.8% (95% CI, 37.5% to 64.1%) to 74.6% (95% CI, 61.6% to 85.0%), with an ROC of 92.8% (95% CI, 88.0% to 97.5%). With this approach, false notifications fell 48%, from 29 to 15, with improvements in sensitivity from 96.9% (95% CI, 89.2% to 99.6%) to 98.4% (95% CI, 91.6% to 100%).

Error Analysis

Error analysis of the expert system, as shown in Table 5, focused on the naïve Bayes algorithm, because this had the best performance characteristics in experiments. We found one missed patient with IMD (false negative) resulting in an overall missed notification rate of 0.8% (one of 123). This patient had a previous lung resection for *Scedosporium apiospermum* infection in an earlier clinical encounter, but the CT scan for the current encounter was appropriately labeled negative by the text classifier because it represented postoperative changes. Although the patient received liposomal amphotericin and voriconazole sequentially, neither exceeded the thresholds to trigger the drug rules for that particular encounter.

There were 15 patients flagged positive by the expert system (false positives) but subsequently not diagnosed with IMD (Table 5). In all cases, the probability of fungal disease according to the text classifier (Scan_Any positive) was well above the prespecified threshold of 0.5. In other words, the baseline text classifier was responsible for the

Table 3. Performance Characteristics of the Classifiers Compared With the Baseline Text Classifier

Algorithm	Threshold*	Sensitivity, % (95% CI)	Specificity, % (95% CI)	ROC, % (95% CI)
Baseline text classifier	0.00	96.9 (89.2 to 99.6)	50.8 (37.5 to 64.1)	73.9 (67.1 to 80.6)
Naïve Bayes	0.32	98.4 (91.6 to 100.0)	74.6 (61.6 to 85.0)	92.8 (88.0 to 97.5)
Random forest	0.12	96.9 (89.2 to 99.6)	72.9 (59.7 to 83.6)	94.1 (89.8 to 98.3)
Logistic	0.00	98.4 (91.6 to 100.0)	50.8 (37.5 to 64.1)	91.1 (85.7 to 96.5)
SMO	0.12	96.9 (89.2 to 99.6)	69.5 (56.1 to 80.8)	92.4 (87.5 to 97.3)
J48	0.10	96.9 (89.2 to 99.6)	57.6 (44.1 to 70.4)	87.2 (80.6 to 93.7)

Abbreviations: ROC, receiver operating characteristic; SMO, sequential minimal optimization.

*The algorithm generated a threshold automatically to maintain sensitivity at least as high as the baseline text classifier while maximizing specificity.

Table 4. Characteristics of Patients With and Without IMD

Characteristic	IMD Group	Control Group
No. of patients	64 (52)	59 (48)
Male sex	39 (61)	27 (46)
Mean age, years (range)	53 (24-89)	51 (18-89)
Underlying disease		
AML	24 (38)	30 (51)
ALL	10 (16)	12 (20)
Lymphoma	14 (22)	7 (12)
Chronic leukemia	6 (9.4)	1 (1.7)
MDS/transformed MDS	6 (9.4)	2 (3.4)
Multiple myeloma	2 (3.1)	3 (5.1)
Other	2 (3.1)	4 (6.8)
Neutropenia (≤ 0.5 cells/L) present	55 (86)	42 (71)
Median duration of neutropenia, days (IQR)	17 (13-26)	18 (10-25)
HSCT	34 (53)	27 (46)
Allogeneic	29 of 34 (85)	21 of 27 (78)
Autologous	5 of 34 (15)	6 of 27 (22)
Characteristic of IMDs (n = 64)		NA
Probable and/or proven IMDs	26 (41)	
Possible IMDs	38 (59)	
Site of infection		
Lung	54 (84)	
Sinopulmonary	3 (4.8)	
Sinus	1 (1.6)	
Hepatosplenic	2 (3.2)	
Disseminated	4 (6.3)	
Organism		
Aspergillus fumigatus	9	
Nonfumigatus *Aspergillus* species (*A. niger*, *A. flavus*)	2	
Fungal hyphae resembling *Aspergillus* species	3	
Scedosporium species	4	
Any positive PCR	4	
Rhizopus species	3	
Other molds (*Acrophialophora fusispora*, *Paecilomyces lilacinus*)	2	
Candida glabrata (coinfection with *Scedosporium prolificans* fungemia)	1	

NOTE. Data presented as No. (%), except as otherwise indicated.

Abbreviations: ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; HSCT, hematopoietic stem-cell transplantation; IMD, invasive mold disease; IQR, interquartile range; MDS, myelodysplastic syndrome; NA, not applicable; PCR, polymerase chain reaction.

majority of false positives; a drug rule was positive in nine encounters, and one encounter had a positive microbiology indicator being Non_Sterile_Any positive (ie, *Aspergillus fumigatus* isolation from sputum) in a patient with multiple myeloma who did not have an IMD.

DISCUSSION

Our expert system is a promising model for electronic surveillance of IMD in hospitals that extends earlier efforts to detect IMD from chest CT reports using NLP.[11,12] High sensitivity maximized case finding, and we reduced false notifications (ie, improved specificity) by combining microbiology and antifungal drug–dispensing data with NLP of CT scan reports.[11,12] Compared with the baseline text classifier, the best performing expert system improved sensitivity slightly from 96.9% to 98.4%. The greatest benefit was seen in specificity, which increased from 50.8% to 74.6%, reducing false notifications by 48% from 29 to 15. Overall accuracy was good, as reflected by area under the ROC, which increased from 73.9% to 92.8% compared with the baseline text classifier alone.

The use of NLP was motivated by the poor sensitivity of other screening approaches for epidemiologic surveillance of IMD. Laboratory-based surveillance of IMD that uses culture and histology is insensitive because microbiology for *Aspergillus* and hyaline molds is positive in ≤ 50% of cases,[17] and patients are often too unwell to undergo invasive diagnostic procedures. Molecular biomarkers, including GM or PCR, are not widely available and have a suboptimal sensitivity,[18,19] which is further reduced by concomitant antifungal therapy administered at the earliest clinical suspicion of IMD.[20] Coding data are unreliable for IMD surveillance[21] and neither timely nor informative enough for outbreak detection. Our expert system that incorporates NLP of CT reports could overcome many of these barriers. It leverages routinely available electronic clinical and operational data that will become easier to access in the computational environment of the electronic medical record (EMR); of note, this work is not dependent on the EMR, which was not present in the study hospitals.

Recent approaches to electronic surveillance of a range of diseases from hospitals have shown promise. A major study demonstrated superiority of NLP of the EMR for surveillance of postoperative complications compared with the more widely used diagnostic codes (eg, sensitivity 82% *v* 38% for acute renal failure).[22] NLP was recognized to have the added benefit of timeliness and the potential to detect while the patient is still in the hospital,[22,23] a characteristic also relevant to IMD because of the risk of hospital acquisition and/or outbreaks associated with these infections.[24]

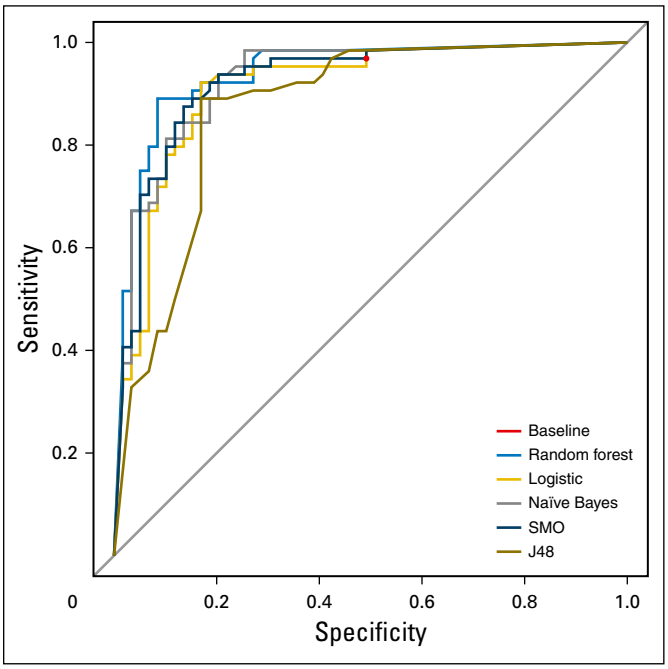

Fig 2. Receiver operating curve of the final classifiers processing probabilistic output from natural language processing of computed tomography reports by the baseline text classifier, microbiology, and antifungal drug dispensing. The red dot represents performance of the baseline text classifier, which had a fixed prespecified threshold of 0.5[11,12] and a relatively high sensitivity without the addition of microbiology and antifungal drug–dispensing data. SMO, sequential minimal optimization.

The quality of source data is another challenge of electronic surveillance. Antifungal drug–dispensing data were least valuable, partially because of their format (pharmacy dispensing is not equivalent to bedside electronic prescribing), and they lacked the specific reasons for drug choice (ie, prophylaxis, empirical, pre-emptive, or targeted therapy). An audit of a computerized decision support system (CDSS) for antifungal drug prescriptions in pediatric hematology-oncology patients demonstrated the superiority of CDSS for inpatient fungal surveillance over coding data,[25] but its usefulness is entirely dependent on prescriber adherence (ie, where prescribers enter their justification for drug choice) with the CDSS.

Other approaches to disease surveillance have combined diagnostic codes with NLP.[26,27] For hepatocellular carcinoma, diagnostic codes were used as a first level of identification, with NLP subsequently applied to pathology and radiology reports to improve case ascertainment.[27] For pneumonia, DeLisle et al[28] combined a text classifier with EMR-derived structured data that included clinical features and diagnostic codes. The addition of text classification of chest radiograph reports increased the positive predictive value of EMR-based case-detection algorithms by 20% to 70%, while retaining sensitivities of 58% to 75%.[28] Ahmed et al[29] used a combination of search term queries of an EMR along with extraction of vital signs, medications, and laboratory values to identify a variety of acute conditions including sepsis, pneumonia, aspiration, acute pancreatitis, and shock, which are risk factors for acute respiratory distress syndrome. For sepsis, they observed that an automated algorithm that combined NLP, vital signs, and laboratory values had a higher sensitivity than NLP alone, which was used in another study (95% *v* 88%).[30] In common among these studies is the improved disease prediction by combining multiple sources of data (structured and unstructured) with NLP, and, indeed, our study supports this approach.

Our study has several limitations. The modest data set of 64 case patients was a major limitation, but emblematic of the difficulties acquiring large training data sets for a disease for which routine epidemiologic surveillance does not exist. The probabilities from the text classifier were derived from our earlier work[11] rather than acquired in a pipeline approach with sequential classifiers operating; however, this should not have appreciably changed our overall findings. Our case patients did not exclusively have proven or probable IMD (despite the fact that these represented a higher degree of certainty) for several reasons: possible IMD cases (those with radiologic features but lacking positive microbiology) represent a substantial burden in clinical practice and may predominate (up to 90%) in some centers[2,31]; possible IMD cases are treated similarly to proven and/or probable IMD cases and consume equivalent health care resources (eg, diagnostics, antifungal drug use),[1] and their exclusion would underestimate the true prevalence of IMD. Notably, all case patients underwent expert adjudication in accordance with international definitions.[14] We focused on the highest-risk population, and limited generalizability to other risk groups (eg, solid organ transplant). Finally, our data set was enriched with positive cases, but external prospective validation of the expert system in clinical practice where prevalence rates are lower, and among diverse patient groups, will be the focus of future work.

A disease with an evolving epidemiology associated with the introduction of novel cancer therapeutics, expanded transplantation services, an aging population, and neglected groups such as children, requires a surveillance system with the agility to keep pace with change. An electronic surveillance system could meet this demand, particularly with the increasing digitization of health care. Our expert system using the wealth of electronic data available in hospitals could be used to flag patients with IMD for further review. The addition of other data analytics to our

Table 5. Error Log of False Negatives and False Positives for Invasive Mold Diseases From the Expert System Incorporating the Naïve Bayes Algorithm

	Output of Text Classifier	Microbiology Investigation							Antifungal Drug Rule[1]			Prediction by Expert System
Patient	Scan_Any positive	Blood_Any positive	BAL_Any positive	Fluid_Any positive	PCR_Any positive	GM_Any positive	Sterile_Any positive	Non_Sterile_Any positive	Drug 1	Drug 2	Drug 3	
False negatives (ie, fungal cases classified as negative)												
1	0	No test	No test	No test	No test	No test	No test	No test	low	low	No	0
False positives (ie, negative cases classified as fungal cases)												
1	0.71	No test	No test	No test	No test	No test	No test	No test	Yes	No	No	0.56
2	0.83	Negative	Negative	Negative	No test	No test	No test	Negative	No	No	No	0.81
3	0.86	No test	No test	No test	No test	No test	No test	No test	Yes	No	No	0.74
4	0.71	Negative	No test	No test	Negative	Negative	No test	positive	Yes	No	No	1.00
5	1.00	Negative	No test	No test	No test	No test	No test	No test	Low	No	No	0.71
6	0.75	Negative	No test	No test	No test	No test	No test	No test	Yes	No	No	0.66
7	0.67	Negative	Negative	No test	No test	No test	No test	No test	Low	Low	No	0.77
8	1.00	No test	No test	No test	No test	No test	No test	No test	No	No	No	0.39
9	1.00	Negative	Negative	No test	Negative	No test	No test	No test	Yes	Low	No	0.98
10	0.75	Negative	Negative	No test	No test	No test	No test	No test	Yes	No	No	0.83
11	1.00	Negative	No test	No test	No test	No test	No test	No test	Low	No	No	0.73
12	0.71	Negative	No test	No test	No test	No test	No test	No test	Low	No	No	0.43
13	0.80	No test	No test	No test	No test	No test	No test	No test	Yes	No	No	0.69
14	0.67	No test	No test	No test	No test	No test	No test	No test	Yes	Low	Low	0.82
15	0.67	Negative	No test	No test	No test	No test	No test	No test	Yes	No	No	0.41

Abbreviations: BAL, bronchoalveolar lavage; GM, galactomannan; PCR, polymerase chain reaction.

expert system, such as image analysis of CT scans, is an area of active research[32] that we hope will improve the performance of the expert system and provide decision support for a range of activities, including radiologic diagnosis and antifungal stewardship programs in hospitals.[8]

DOI: https://doi.org/10.1200/CCI.17.00011
Published online on ascopubs.org/journal/cci on August 30, 2017.

AUTHOR CONTRIBUTIONS

Conception and design: Michelle R. Ananda-Rajah, Christoph Bergmeir, Francois Petitjean, Monica A. Slavin, Karin A. Thursky, Geoffrey I. Webb
Administrative support: Michelle R. Ananda-Rajah
Provision of study material or patients: Michelle R. Ananda-Rajah, Monica A. Slavin, Karin A. Thursky
Collection and assembly of data: Michelle R. Ananda-Rajah
Data analysis and interpretation: All authors
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

Michelle R. Ananda-Rajah
Honoraria: Gilead Sciences, Merck Sharp & Dohme
Research Funding: Merck Sharp & Dohme (Inst)
Travel, Accommodations, Expenses: Gilead Sciences

Christoph Bergmeir
No relationship to disclose

François Petitjean
Stock and Other Ownership Interests: Vanguard VTS and VEU indexes

Monica A. Slavin
Travel, Accommodations, Expenses: Gilead Sciences

Karin A. Thursky
No relationship to disclose

Geoffrey I. Webb
No relationship to disclose

ACKNOWLEDGMENT

We thank Dr Orla Morrissey for reviewing the manuscript.

REFERENCES

1. Ananda-Rajah MR, Cheng A, Morrissey CO, et al: Attributable hospital cost and antifungal treatment of invasive fungal diseases in high-risk hematology patients: An economic modeling approach. Antimicrob Agents Chemother 55:1953-1960, 2011
2. Ananda-Rajah MR, Grigg A, Downey MT, et al: Comparative clinical effectiveness of prophylactic voriconazole/posaconazole to fluconazole/itraconazole in patients with acute myeloid leukemia/myelodysplastic syndrome undergoing cytotoxic chemotherapy over a 12-year period. Haematologica 97:459-463, 2012
3. Kontoyiannis DP, Marr KA, Park BJ, et al: Prospective surveillance for invasive fungal infections in hematopoietic stem cell transplant recipients, 2001-2006: Overview of the Transplant-Associated Infection Surveillance Network (TRANSNET) Database. Clin Infect Dis 50:1091-1100, 2010
4. Juliusson G, Antunovic P, Derolf A, et al: Age and acute myeloid leukemia: Real world data on decision to treat and outcomes from the Swedish Acute Leukemia Registry. Blood 113:4179-4187, 2009
5. Lortholary O, Gangneux JP, Sitbon K, et al: Epidemiological trends in invasive aspergillosis in France: The SAIF network (2005-2007). Clin Microbiol Infect 17:1882-1889, 2011
6. Pappas PG, Alexander BD, Andes DR, et al: Invasive fungal infections among organ transplant recipients: Results of the Transplant-Associated Infection Surveillance Network (TRANSNET). Clin Infect Dis 50:1101-1111, 2010
7. Even C, Bastuji-Garin S, Hicheri Y, et al: Impact of invasive fungal disease on the chemotherapy schedule and event-free survival in acute leukemia patients who survived fungal disease: A case-control study. Haematologica 96:337-341, 2011
8. Ananda-Rajah MR, Slavin MA, Thursky KT: The case for antifungal stewardship. Curr Opin Infect Dis 25:107-115, 2012
9. Valerio M, Muñoz P, Rodríguez CG, et al: Antifungal stewardship in a tertiary-care institution: A bedside intervention. Clin Microbiol Infect 21:492.e1-9, 2015
10. Ananda-Rajah MR: Re-imagining the surveillance of invasive mould diseases [doctoral thesis]. The Royal Melbourne Hospital, Melbourne, Victoria, Australia, 2014
11. Ananda-Rajah MR, Martinez D, Slavin MA, et al: Facilitating surveillance of pulmonary invasive mold diseases in patients with haematological malignancies by screening computed tomography reports using natural language processing. PLoS One 9:e107797, 2014

12. Martinez D, Ananda-Rajah MR, Suominen H, et al: Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans. J Biomed Inform 53:251-260, 2015
13. Nicolle MC, Bénet T, Thiebaut A, et al: Invasive aspergillosis in patients with hematologic malignancies: Incidence and description of 127 cases enrolled in a single institution prospective survey from 2004 to 2009. Haematologica 96: 1685-1691, 2011
14. De Pauw B, Walsh TJ, Donnelly JP, et al: Revised definitions of invasive fungal disease from the European Organization for Research and Treatment of Cancer/Invasive Fungal Infections Cooperative Group and the National Institute of Allergy and Infectious Diseases Mycoses Study Group (EORTC/MSG) Consensus Group. Clin Infect Dis 46:1813-1821, 2008
15. Patterson TF, Thompson GR III, Denning DW, et al: Practice guidelines for the diagnosis and management of aspergillosis: 2016 update by the Infectious Diseases Society of America. Clin Infect Dis 63:e1-e60, 2016
16. Hornik K, Buchta C, Zeileis A: Open-source machine learning: R meets Weka. Comput Stat 24:225-232, 2009
17. Denning DW: Early diagnosis of invasive aspergillosis. Lancet 355:423-424, 2000
18. Mengoli C, Cruciani M, Barnes RA, et al: Use of PCR for diagnosis of invasive aspergillosis: Systematic review and meta-analysis. Lancet Infect Dis 9:89-96, 2009
19. Pfeiffer CD, Fine JP, Safdar N: Diagnosis of invasive aspergillosis using a galactomannan assay: A meta-analysis. Clin Infect Dis 42:1417-1427, 2006
20. Maertens J, Groll AH, Cordonnier C, et al: Treatment and timing in invasive mould disease. J Antimicrob Chemother 66:i37-i43, 2011 (suppl 1)
21. Chang DC, Burwell LA, Lyon GM, et al: Comparison of the use of administrative data and an active system for surveillance of invasive aspergillosis. Infect Control Hosp Epidemiol 29:25-30, 2008
22. Murff HJ, FitzHenry F, Matheny ME, et al: Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 306:848-855, 2011
23. Webb KH: Natural language processing and electronic medical records. JAMA 306:2325-2326, 2011
24. Tomblyn M, Chiller T, Einsele H, et al: Guidelines for preventing infectious complications among hematopoietic cell transplantation recipients: A global perspective. Biol Blood Marrow Transplant 15:1143-1238, 2009
25. Bartlett A, Goeman E, Vedi A, et al: A pilot study of a computerized decision support system to detect invasive fungal infection in pediatric hematology/oncology patients. Infect Control Hosp Epidemiol 36:1313-1317, 2015
26. Danforth KN, Early MI, Ngan S, et al: Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. J Thorac Oncol 7:1257-1262, 2012
27. Sada Y, Hou J, Richardson P, et al: Validation of case finding algorithms for hepatocellular cancer from administrative data and electronic health records using natural language processing. Med Care 54:e9-e14, 2016
28. DeLisle S, Kim B, Deepak J, et al: Using the electronic medical record to identify community-acquired pneumonia: Toward a replicable automated strategy. PLoS One 8:e70944, 2013
29. Ahmed A, Thongprayoon C, Pickering BW, et al: Towards prevention of acute syndromes: Electronic identification of at-risk patients during hospital admission. Appl Clin Inform 5:58-72, 2014
30. FitzHenry F, Murff HJ, Matheny ME, et al: Exploring the frontier of electronic health record surveillance: The case of postoperative complications. Med Care 51:509-516, 2013
31. Neofytos D, Treadway S, Ostrander D, et al: Epidemiology, outcomes and mortality predictors of invasive mold infections among transplant recipients: A 10-year, single-center experience. Transpl Infect Dis 15:233-242, 2013
32. Monash University: Monash Institute of Medical Engineering, 2015 Seed Fund Results. https://www.monash.edu/mime/seed/2015-results