Natural Language Processing and Machine Learning to Enable Clinical Decision Support for Treatment of Pediatric Pneumonia

Joshua C. Smith, PhD, Ashley Spann, MD, Allison B. McCoy, PhD,
Jakobi A. Johnson, BS, Donald H. Arnold, MD, MPH,
Derek J. Williams, MD, MPH, Asli O. Weitkamp, PhD
Vanderbilt University Medical Center, Nashville, TN

Abstract

Pneumonia is the most frequent cause of infectious disease-related deaths in children worldwide. Clinical decision support (CDS) applications can guide appropriate treatment, but the system must first recognize the appropriate diagnosis. To enable CDS for pediatric pneumonia, we developed an algorithm integrating natural language processing (NLP) and random forest classifiers to identify potential pediatric pneumonia from radiology reports. We deployed the algorithm in the EHR of a large children's hospital using real-time NLP. We describe the development and deployment of the algorithm, and evaluate our approach using 9-months of data gathered while the system was in use. Our model, trained on individual radiology reports, had an AUC of 0.954. The intervention, evaluated on patient encounters that could include multiple radiology reports, achieved a sensitivity, specificity, and positive predictive value of 0.899, 0.949, and 0.781, respectively.

Introduction

Integration of natural language processing (NLP) and machine learning to guide real-time decision support could allow for streamlined clinical care but is overall lacking in clinical practice, particularly in the case of pediatric pneumonia.[1] Recently, a clinical trial at our institution employed a clinical decision support (CDS) strategy to promote appropriate antibiotic prescribing for children with pneumonia presenting to the Emergency Department at Monroe Carell Jr. Children's Hospital at Vanderbilt.[2,3] This CDS was triggered when pneumonia was added to a pediatric patient's problem list in the EHR. To facilitate study enrollment, it was necessary to help providers recognize potential pneumonia as soon as possible, and prompt them to add the diagnosis to the problem list. To that end, we developed a novel method for identifying chest x-ray (CXR) radiology reports that support a diagnosis of pediatric pneumonia. This paper describes the development of this method, which uses NLP on radiology reports combined with a random forest classifier, and an evaluation based on 9-months of data gathered while the system was in use.

Background

Pneumonia is an acute respiratory infection affecting the alveoli and distal bronchioles. It is the most frequent cause of infectious disease-related deaths in children worldwide, accounting for 15% of deaths in those less than 5 years of age in 2017.[4,5] In a cross-sectional study of U.S. acute care hospitals, pneumonia was the most frequent and expensive reason for pediatric hospitalizations.[6] Community-acquired pneumonia (CAP), defined as pneumonia acquired outside a hospital setting, accounts for approximately two million outpatient visits and 124,000 pediatric hospitalizations annually.[5] Causative organisms include mainly viruses and bacteria.[7] Complications include parapneumonic effusion, empyema, necrotizing pneumonia, pulmonary abscess, acute respiratory failure and sepsis. Signs and symptoms of pneumonia include fever, hypoxemia, tachypnea, cough, chest pain, dyspnea and grunting; physical examination findings include rales, wheezing, and dullness to auscultation. Many of these signs and symptoms are nonspecific and may result from atelectasis, acute asthma exacerbations, bronchiolitis, congestive heart failure, and other etiologies.

Chest radiography is the most frequently ordered diagnostic test for suspected pneumonia.[8] Yet, as with signs, symptoms, and physical exam findings, CXR findings among children with pneumonia may be nonspecific and experts often disagree on the interpretation.[9,10] A systematic review of children with radiographic evidence of pneumonia found that most signs and symptoms have limited predictive value, and there is considerable overlap of these findings in pneumonia caused by different etiologies.[11] As there is no criterion standard for CXR diagnosis of pediatric pneumonia, and CXR cannot distinguish between viral and bacterial infections, diagnosis and treatment decisions must be made in the context of clinical, laboratory, and radiologic findings.[6] Notwithstanding the limitations of CXR, there is a spectrum of radiological appearances consistent with the clinical and pathological diagnosis of pneumonia and findings may expedite appropriate clinical management and improve outcomes.[8] As a result, CXR has become the gold standard for confirmation of a clinical diagnosis of pneumonia.

A number of previous studies have attempted to recognize pneumonia from CXR reports and other clinical data, with significant early research being performed at the University of Utah and LDS Hospital in Salt Lake City.[12–18] In 1994, Haug, et al., developed an early "natural language understanding" system to parse full text CXR reports to identify important findings and map them to a controlled medical vocabulary, but the system proved difficult to maintain at the time.[12] In 1999, Chapman & Haug found that computerized techniques can perform as well as a physician, concluding that machine learning can be used to identify CXR reports that support a pneumonia diagnosis.[13] Also in 1999, Aronsky & Haug developed a Bayesian Network to identify adult pneumonia in the emergency department (ED).[14] The inputs for the Bayesian Network included a number of demographic and clinical variables, triage information, breath sounds, and lab values, as well as coded data recorded by the ED physician after viewing the completed CXR report; codes indicated the specific presence of pneumonia and/or effusion. They recognized, however, that CDS systems that rely on additional and time-consuming manual data entry result in a "behavioral bottleneck" that prevents such systems from becoming part of clinical routine.[14] Their system achieved an AUC of 0.930 and, at a fixed sensitivity of 0.95, specificity and PPV of 0.685 and 0.073, respectively; when the sensitivity was fixed at 0.90, specificity and PPV were 0.790 and 0.102, respectively.[15]

In 2000, Fiszman, et al., compared SymText, an early NLP system, and keyword searches of CXR reports, combined with expert-crafted rules, against physicians; they found these approaches compared favorably with physicians, achieving sensitivity, specificity, and PPV of 0.95, 0.85, and 0.78, respectively.[16] However, their analysis was performed only on a small, manually-curated gold-standard set of 292 reports enriched to include additional examples of bacterial CAP, and results were admittedly unlikely to generalize to reports representing actual disease prevalence. In 2001, Aronsky, et al., combined their Bayesian Network with SymText-identified keywords to recognize pneumonia from adult patients in the ED.[17] However, they found that the physician-coded data on pneumonia and effusions, important for their Bayesian Network, was often missing; NLP seemed to compensate for this missing data, and resulted in improved performance over previous efforts. Lagor, et al., compared the Bayesian Network approach with Artificial Neural Networks, trained on over 32,000 patients, and found similar results.[18] Their best-performing algorithm achieved a specificity of 0.940 and PPV of 0.186 at a fixed 0.95 sensitivity, but both methods included 25 manually-curated variables, including, breath sounds, chief complaints, and a number of other clinical characteristics.

Other research has focused on different populations, clinical documents, and techniques. Liu, et al., studied the use of NLP on intensive care unit (ICU) reports and reported success in identifying pneumonia in critically ill patients.[19] In 2008, Elkin, et al., used NLP on CXR reports to obtain SNOMED-CT codes. After applying expert crafted rules, they concluded that SNOMED-CT-based rules and codes were accurate enough to detect pneumonia from CXR reports.[20] Asatryan, et al. again focused on using NLP to identify keywords and found that "pneumonia" was most predictive, as might have been suspected.[21] However, simple keywords can be misleading, as CXR reports often include a significant amount of "hedging," revealing uncertainty about the diagnosis.[22]

Most of these previous efforts were focused on adult patient populations, but the presentation of pneumonia in children can differ.[8] Presentation of pneumonia can also differ based upon the pathogen and type of pneumonia.[23] In 2005, Mendonça, et al., utilized NLP on CXR reports to recognize hospital-acquired pneumonia in neonates, achieving a sensitivity of 0.71 and specificity of 0.99, but a PPV of only 0.075.[24] In 2017, Meystre, et al, used these approaches to diagnose bacterial pneumonia in a pediatric patients.[25] They used NLP to extract specific findings from CXR reports for input to a Support Vector Machine (SVM). They achieved sensitivity, specificity, and PPV of 0.71, 0.96, and 0.86, respectively, often performing better than domain experts. The system was developed using a manually-curated gold standard of 282 reports to extract information from a larger collection of CXR reports, but was not real-time.

Summary

This paper describes the development and evaluation of a method that utilizes NLP and random forests to identify pediatric pneumonia from radiology reports. Unlike the majority of previous work, it does not incorporate clinical variables imported from other parts of the EHR or manual input by physicians. Manually curated clinical data, especially pneumonia-specific variables used in previous work, can improve prediction, but are often delayed or absent due to the aforementioned behavioral bottleneck.[14] This intervention, supporting rapid initial recognition of possible pneumonia in a general pediatric population, was implemented in a large children's hospital using real-time NLP. We analyzed performance based on 9-months of real clinical data gathered while the system was in use.

Methods

To identify radiology reports that support diagnosis of pediatric pneumonia, we utilized a random forest classifier using NLP-extracted features as input. A random forest is a machine-learning classifier based on an ensemble of

decision trees.[26] It is trained using a corpus of cases and controls. We developed our training corpus by first identifying approximately 10,000 historical pediatric admissions at our institution. Cases were a subpopulation of a larger cohort of the Etiology of Pneumonia in the Community (EPIC) study, who were enrolled at Monroe Carell Jr. Children's Hospital at Vanderbilt.[2,7] Patients were enrolled in the study if they were under 18 years of age, hospitalized with signs or symptoms of acute infection and acute respiratory illness, and had clinical and radiographic evidence of pneumonia. Children with recent hospitalization, severe immunosuppression, cystic fibrosis, tracheostomy, or clear alternative diagnosis were excluded. Controls were identified based upon availability of CXR and the absence of pneumonia-related diagnostic codes. The initial training corpus consisted of approximately 10% cases and 90% controls.

For each of the 10,000 admissions, we queried an EHR data repository to extract the earliest radiology report available within 24 hours of admission, either before or after. Not all patients in the cohort had radiology reports in their record within the specified window. Additionally, a number of these patients had CXR performed at outside institutions and, if the associated reports was in our EHR, it was saved as a scanned image or PDF and was not used in the study. This resulted in 5053 CXR reports (4314 controls and 739 cases). While all reports had a separate *Impressions* section, the rest of the report was either in a single, unlabeled section, or separated into individual sections such as, *Indication, Findings, History,* and *Comparisons*; in all cases, the complete report was utilized.

After removing any XML markup from the report text, we processed the resulting plain text files using *MetaMapLite*.[27] MetaMap is an NLP tool that identifies Unified Medical Language System (UMLS) concepts mentioned in clinical text; it incorporates the NegEx algorithm to identify whether mentioned concepts are negated (i.e., "opacity in right lung" vs "no opacities").[28,29] MetaMapLite provides similar functionality with reduced overhead and greater speed.[27] For negation detection, it can use either NegEx or the more robust ConText[30] algorithm; we opted to use NegEx because it is slightly faster. We also restricted concepts to the SNOMED-CT vocabulary (validated for pneumonia by Elkin, et al.[20]) and represented using UMLS Concept Unique Identifiers (CUIs) from the UMLS 2017AA release.

We processed the MetaMapLite output to extract mentioned concepts and converted each report into a set of CUIs with their negation status (i.e., "affirmed_C0032285", "negated_C0034063"). We then identified the most frequent CUI/negation-status pairs (those occurring at least five times) from among the entire training corpus. Using the resultant list of 1021 frequent concepts, each report in the training corpus was transformed into a binary vector representing the presence or absence of these affirmed/negated CUIs. These vectors were then used as input for a random forest classifier using R (version 3.5.0) and the R package *randomForest*.[31,32] We constructed our classifiers using 700 decision tress and the package default values for other parameters (*mtry, nodesize, maxnodes*, etc.). Due to computational limitations, we did not perform a thorough grid search to optimize hyperparameters; experimenting with additional trees and different values of *mtry* did not improve performance.

To validate our training data, we built an initial random forest model using all available cases and controls. That process identified a number of low-scoring *cases,* which the model classified as likely having no pneumonia, and high-scoring *controls,* which the model classified as supporting pneumonia diagnosis. Physicians reviewed these approximately 200 radiology reports and, as expected, many had been misclassified originally. Flagged reports were then reclassified (cases to controls, and vice versa) resulting in 823 cases and 4230 controls.

After this data validation step, we then rebuilt our random forest model using 10-fold cross validation, training and testing each iteration of the model on approximately 4500 and 500 reports, respectively. We also constructed a model using the entire training corpus (with no test set). We report the area under the receiver operating characteristic curve (AUC) on each test set, as well as the AUC calculated on the out-of-bag sample (OOB AUC) for the training sets. The OOB AUC provides an estimate of the model's performance based on unused portions of the training data. We then reviewed the expected performance of the model at various thresholds and, based on desired precision and recall for the aforementioned clinical trial, determined the specific decision threshold to be used.

Implementation

After appropriate testing and validation, the NLP process and random forest model were deployed as part of the CDS infrastructure and integrated with our Epic® EHR. The decision support infrastructure at our institution utilizes an Enterprise Service Bus (ESB) that serves as an integration platform for all inbound and outbound data flow associated with clinical systems. Via specific adaptors, the ESB integrates CDS services with clinical systems for real-time data processing; clinical events represented by discrete data flowing through the ESB initiate surveillance and update CDS service inputs/outputs in real-time. In this case, when a patient was admitted to the Pediatric ED, the system began listening for CXR orders for the next 24-hours. As soon as they were available on the ESB, the system then processed and scored any corresponding radiology reports. If any report scored above the decision threshold (i.e., potential

pneumonia), the system posted a smart data element using Epic's APIs to flag the patient. The presence of the smart data element triggered an interruptive Best Practice Advisory (BPA) alert informing the providers that, based upon the CXR report, the patient was suspected of having pneumonia. The BPA displayed the radiology report impression section and allowed the provider to (1) add CAP to the problem list, (2) report that the patient did not have pneumonia, or (3) simply acknowledge or defer the alert. Providers could also enter optional comments. If the provider added CAP to the problem list, this would fire a second BPA describing the clinical trial, including any exclusion criteria (age less than six months, age greater than 18 years, cystic fibrosis, immunosuppression, or hospitalized within the last seven days). This workflow allowed the provider to either enroll the patient, which would trigger a CAP Antibiotic Adviser, or decline and note whether their exclusion from the trial was contraindicated or due to other clinical judgement. If provider did not add CAP to the problem list, declined to enroll in the trial, or reported no pneumonia, no further alerts were generated. Due to the minimal-risk nature of the clinical trial, enrollment decisions were made by the provider.

Evaluation Methods

When a CXR is ordered, the order frequently includes multiple images, or views, and each image has its own associated report. Sometimes the reports in the set are identical; other times they are different and pneumonia is only recognizable in some views/reports. If any of the individual reports were flagged by the NLP algorithm, it triggered the BPA. In other words, the algorithm is based upon sets of reports from the same session or encounter, rather than individual CXR reports, and we evaluated it accordingly. We considered an encounter as positive for potential pneumonia if any report from the encounter was flagged as potential pneumonia. Correspondingly, we considered the encounter negative for potential pneumonia when no reports from the encounter were flagged by the algorithm. We evaluated the performance of the algorithm using 9-months of data collected from January 1 to October 1, 2019.

The algorithm classified all encounters with CXR as either "radiology reports support a diagnosis of pneumonia," denoted as NLP-positive (NLP+) or "radiology reports do not support a diagnosis of pneumonia," denoted as NLP-negative (NLP-). It is important to note that pneumonia is not solely a radiological diagnosis. We were not evaluating the sufficiency of CXR reports to diagnose pneumonia, but rather whether the algorithm was able to identify those radiology reports that do support a diagnosis of possible pneumonia. Further, while the trial is focused solely on identifying CAP, the algorithm was designed to identify CXR reports that support any possible pneumonia diagnosis, and we evaluated it accordingly. Reviewing all CXR reports in the study was not feasible due to the large number of reports. To identify which reports required manual review to resolve true and false positives and negatives, we considered provider responses to the BPA, patient problem lists, and encounter diagnoses. In cases with conflicting information (e.g., problem lists or encounter diagnosis indicated pneumonia but the algorithm did not, and vice versa), physician reviewers read radiology reports to identify whether any report in a given encounter "supported a diagnosis of possible pneumonia."

Among NLP+ encounters, the provider either (1) added CAP to the problem list from the BPA or (2) did not add the problem using the BPA. Among the latter group, there were a minority of instances where the provider never received the BPA because the patient had been discharged from the ED or admitted to another service. To resolve true and false *positives*, we manually reviewed these and all other CXR reports from NLP+ encounters where the provider did not add the problem using the BPA (including encounters where patients met exclusion criteria for the clinical trial). We reviewed reports from encounters where the provider did add CAP only if that problem was subsequently deleted from the problem list. To resolve true and false *negatives*, we reviewed reports from NLP- encounters when there was any evidence of any pneumonia in the patient's problem list or encounter diagnoses. We did not review NLP- encounters if there was no evidence of a pneumonia diagnosis.

We report sensitivity, specificity, and PPV of the algorithm, among other measures, and compare to the performance of previous work and the estimated model performance from training data. We also review the most important features of our random forest model as determined by the mean decrease in Gini coefficient.[32,33] Finally, we review provider comments from the BPA and highlight report characteristics that may have resulted in misclassification.

Results

Model Development and Determining a Decision Threshold

The performance of the random forest models on both the training set (OOB AUC) and test sets (AUC) is shown in Table 1. The average AUC across all training sets from the cross-validation was 0.953; the average AUC across the test sets was 0.952. The OOB AUC on the model utilizing all training data was 0.954, revealing that the OOB AUC was similar to using set-aside test sets. This model trained on all available reports was the one we implemented in the EHR. Figure 1 shows the ROC curve of the implemented model.

Table 1. Out-of-bag (OOB) AUC on training sets and AUC on test sets.

Partition	Train (OOB AUC)	Test (AUC)
1	0.955	0.932
2	0.952	0.961
3	0.954	0.933
4	0.951	0.966
5	0.951	0.974
6	0.955	0.941
7	0.951	0.963
8	0.953	0.961
9	0.953	0.951
10	0.956	0.940
Average	0.953	0.952
All Data	**0.954**	

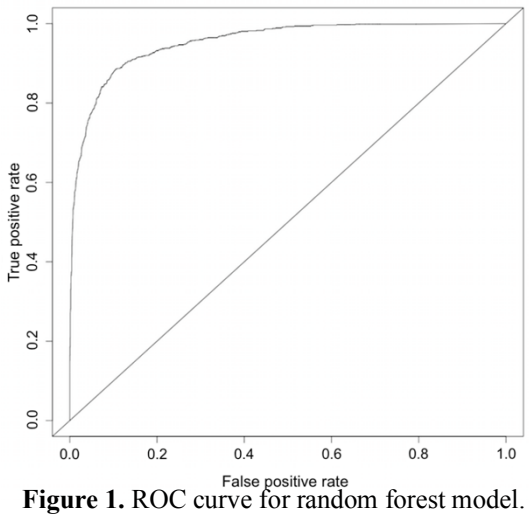

Figure 1. ROC curve for random forest model.

Table 2 illustrates model performance on the training data using various decision thresholds. Measures include sensitivity, specificity, PPV, false positive rate, balanced accuracy ((*sensitivity* + *specificity*)/2), and f-measure. After reviewing the performance on test sets at various thresholds, the clinical team selected a decision threshold of 0.27. It was selected primarily for high sensitivity (88%), acceptable PPV (63%), and a 10% false positive rate.

Table 2. Expected performance of the model at various thresholds; we selected 0.27 as the decision threshold.

Decision Threshold	0.10	0.20	0.27*	0.30	0.40	0.50	0.60	0.70	0.80	0.90
Sensitivity	0.954	0.913	**0.881**	0.860	0.774	0.655	0.482	0.306	0.126	0.002
Specificity	0.729	0.851	**0.899**	0.911	0.948	0.977	0.992	0.999	1.000	1.000
PPV	0.406	0.544	**0.629**	0.653	0.742	0.849	0.925	0.977	0.990	1.000
False Positive Rate	0.271	0.149	**0.101**	0.089	0.052	0.023	0.008	0.001	0.000	0.000
Balanced accuracy	0.841	0.882	**0.890**	0.886	0.861	0.816	0.737	0.652	0.563	0.501
F-measure (F1)	0.570	0.681	**0.734**	0.743	0.757	0.739	0.634	0.466	0.224	0.005

To assess the face validity of the random forest model, we reviewed the most important features as determined by mean decrease in Gini. The most important feature was an affirmative mention of pneumonia (*C0032285*). Other top features included multiple CUIs for opacity (*C1265876, C0029053, C0449584*) and consolidation (*C0521530, C0702116*), specific pneumonia concepts (e.g., *left/right lower/upper/middle lobe pneumonia*, *atypical pneumonia*), and CUIs for *infiltrate, atelectasis,* and *pneumothorax*. Anatomical locations alone (*left/right lower/upper/middle lobe of lung*) and qualifiers such as *patchy, normal, interstitial, focal,* and *multifocal* were also present among the most important features, as well as several symptom concepts (*fever, cough*). While negated features (including *pneumonia*) were present among the most important features, non-negated concepts tended to be ranked higher.

Implementation and Evaluation

During the 9-month study period, the system processed approximately 8600 radiology reports from 3012 distinct patient encounters. The system flagged patients as having suspected pneumonia in 579 encounters; for the remaining 2433 encounters, no radiology reports scored above the decision threshold. Of the 579 NLP+ results, providers added CAP to the patient's problem using the BPA 344 times (68.3%). In 5.3% of encounters, radiology reports included both NLP+ and NLP- reports, and were therefore classified as positive encounters.

For NLP+ encounters where providers added CAP to the problem list through the BPA, the pneumonia problem was deleted soon afterwards in 12 cases (3.5%). Reviewers determined that in 8 of those cases, reports did in fact support a diagnosis of possible pneumonia. In two of those instances, the problem was replaced by other *non-CAP* pneumonia

problems (*pneumonia and influenza* and *aspiration pneumonia*); two others were deleted because there were duplicate CAP entries in the problem list.

For the remaining 235 NLP+ encounters where the provider did not add CAP using the BPA, reviewers found that 48% (112 encounters) had reports that supported a diagnosis of possible pneumonia, and were therefore true positives, and 52% (123 encounters) did not, and were therefore false positives. Among those NLP+ encounters where any pneumonia was included in the problem list or encounter diagnoses, 95% had supporting radiology reports.

Among encounters where the problem was not added, the BPA was triggered 432 times: in 49% of cases, the alert was deferred until further information could be gathered; in 31% of cases, the provider indicated the patient did not have pneumonia; in 13% of cases, the provider responded they were not the attending physician; for the remainder of cases, the alert was acknowledged or overridden without a provided reason. When providers responded that the patient did not have pneumonia, few added specific comments. In two cases, the providers commented that CXR findings likely represented atelectasis or there was no consolidation on x-ray. In four cases, the providers stated that patients did not have CAP, but other pneumonias (hospital acquired, bronchopneumonia, aspiration pneumonia, or viral pneumonia). In three cases, providers indicated there were no clinical signs of pneumonia, and in two more cases, indicated the patient had sickle cell disease or acute chest syndrome. Among all NLP+ encounters when enrollment in the trial was contraindicated, providers still used the BPA to add CAP to the problem list 48 times (86%). Only one provider who added CAP to the problem list through the BPA noted that the it was a clinical, rather than radiological, diagnosis.

For the NLP- encounters, there was evidence of potential pneumonia in the form of problem list entries or encounter diagnoses in only 141/2433 encounters (5.8%). Physicians reviewed all CXR reports from those encounters to identify any false negatives. They found 51 false negatives (36%) where the CXR report did support a diagnosis of possible pneumonia and 90 true negatives (64%) where reports did not support the diagnosis.

Table 3 illustrates the true and false positives and negatives determined by review. Table 4 shows the algorithm's performance during the 9-month study period. Sensitivity, specificity, and PPV were 90%, 95%, and 78%, respectively, compared to 88%, 90%, and 63% on the training set.

Table 3. Confusion matrix illustrating actual algorithm performance. Class refers to encounters with CXR reports that support a pneumonia diagnosis.

		Actual Class		
		Yes	**No**	**Sum**
Predicted Class	**NLP+**	452	127	*579*
	NLP-	51	2382	*2433*
	Sum	*503*	*2509*	

Table 4. Algorithm performance identifying CXR reports that support pneumonia diagnosis.

Sensitivity	0.899
Specificity	0.949
PPV	0.781
False Positive Rate	0.051
Balanced Accuracy	0.924
F-Measure (F1)	0.835

Manual Review of Radiology Reports

To identify false positives alerts, reviewers read all NLP+ radiology reports where CAP was not added to the problem list through the BPA. Among the 127 false positive encounters, only 21 had some form of pneumonia added to the problem list; in the majority of other cases, patients were diagnosed with other respiratory conditions (atelectasis, pleural effusion, bronchiolitis, bronchitis, etc.). In approximately 10% of false positive encounters, patients appeared to have findings of cystic fibrosis, sickle cell anemia, or acute chest syndrome. Reviewers noted that some reports only mentioned pneumonia in an *Indication* section that listed suspected pneumonia as the reason for the CXR order. Many false positive reports mentioned recent pneumonia, a history of pneumonia, or resolved/resolving pneumonia (e.g., "interval resolution of the right apical opacity"). Reports frequently mentioned findings which could represent pneumonia, but the *Impression* sections further specified higher likelihood of other diagnoses, such as atelectasis, bronchitis, bronchiolitis, and peripheral or reactive airway disease. Reviewers noted difficulty with words and phrases denoting uncertainty (e.g., "infection not excluded," "underlying pneumonia difficult to definitively exclude", "pneumonia is considered much less likely", and "opacities favored to represent atelectasis over pneumonia") when other diagnoses were more supported. Finally, reviewers noted that pneumonia was frequently mentioned in specific

terms (e.g., "right lower lobe pneumonia," "right middle lobe pneumonia," "viral pneumonia," "atypical pneumonia," "multifocal pneumonia," bibasilar pneumonia, etc.).

To identify false negatives, reviewers read all CXR reports from NLP- encounters where there was evidence of pneumonia in patients' problem lists or encounter diagnoses, and found 51 reports that supported a possible pneumonia diagnosis. Reviewers noted several common terms from these reports that did not seem to have the desired impact on classification: empyema, airspace opacity, opacification, multifocal, infection/infected, consolidation/consolidative, and air bronchograms. For many false negatives, reviewers noted that pneumonia was mentioned in the impression section, but not the findings section. Additionally, a number of false negatives were in patients with scoliosis, possibly contributing to poor-quality imaging and insufficient information in the report.

Discussion

Using approximately 5000 historical CXR reports from our institution, we developed an algorithm using NLP and a random forest classifier to identify radiology reports that support a diagnosis of pediatric pneumonia. Our classifier had an AUC of 0.954. Based upon 9-months of real clinical data, the sensitivity, specificity, and PPV or our algorithm was 0.899, 0.949, and 0.781, respectively. Developing and analyzing the algorithm in context of a clinical trial was both convenient and beneficial. Domain experts leading and participating in the trial gathered and reviewed initial training data and assisted in reviewing preliminary and final results. Importantly, this work demonstrates that real-time NLP can effectively be integrated into real-world clinical care to improve problem list completeness and delivery of CDS.

The algorithm performed favorably compared to prior work using CXR and other clinical data to recognize adult pneumonia. Our results were significantly better than the sensitivity, specificity, and PPV reported by Aronsky & Haug's Bayesian network approach (0.900, 0.790, 0.102), Aronsky, et al., and their Bayesian Network with NLP (of 0.940 and PPV of 0.186 at a fixed 0.95), and the artificial neural networks used by Lagor, et al. (0.950, 0.940, 0.186).[14,15,17,18] All these methods included a number of manually-curated variables, and the low PPV would likely have caused feasibility issues for CDS. The performance of the Fiszman, et al. approach, combining NLP with expert crafted rules, was more comparable to our method (0.95, 0.85, and 0.78), but was analyzed on only a small, enriched set of CXR reports, and was admittedly unlikely to generalize to the larger patient population.[16]

We identified few similar studies that focused on pediatric populations. The most comparable study (Meystre, et al.), which used SVMs and NLP to diagnose bacterial pneumonia in a general pediatric population, achieved sensitivity, specificity, and PPV of 0.71, 0.96, and 0.86, respectively.[25] Our algorithm achieved slightly lower specificity and PPV, but significantly higher sensitivity (0.899). The fact that our algorithm identifies any type of pediatric pneumonia, rather than strictly bacterial pneumonia, may contribute to performance differences. It is also important to note that their study utilized a small, manually-curated gold standard (a time-intensive process), whereas ours used a sample of convenience from a previous study, partially validated using the initial random forest model. Additionally, their algorithm was applied to a database of historical reports, whereas our algorithm was implemented in the EHR using real-time NLP.

Our algorithm performed better than expected when compared to the training and test data. Prior to implementation, we estimated a sensitivity, specificity, and PPV of 0.881, 0.899, and 0.629, respectively, at the selected decision threshold. In practice, the algorithm achieved sensitivity, specificity, and PPV of 0.899, 0.949, and 0.781. We believe discrepancies are likely due to training and testing using a report-based model and implementing/reviewing the performance using an encounter-based approach. That is, the algorithm analyzed reports from any CXR order in first 24-hours of each encounter; this usually included multiple views, with each image having its own associated radiology report (2.8 images/reports per encounter). Many of the reports in an encounter were identical to one-another, but 5.3% of encounters contained reports with conflicting results (at least one report was NLP+ and one was NLP-). Some of these report-level classifications are likely incorrect, but it is important to realize that pneumonia is not always evident in all CXR views. While they did not exhaustively review all individual reports to evaluate this, reviewers noted cases where different reports (views) from a single encounter supported opposite conclusions. In the other 94.7% of encounters, however, all reports in a given encounter/order set resulted in the same conclusion.

While not the primary focus of this paper, it is important to note the usefulness of the NLP/random forest algorithm in helping enroll patients in the CAP clinical trial. The BPA alerts triggered by the algorithm facilitated the enrollment of 344 patients into the trial. Some clinicians made treatment decisions early in the ED encounter, which often prevented their patients from being eligible for the trial. While most clinicians waited for returned CXR reports to make decisions, the algorithm proved critical in encouraging providers to add CAP to the problem list before starting treatment. Adding CAP to the problem list then triggered the BPA for enrollment and the CAP Antibiotic Advisor. The real-time nature of the NLP enabled nearly instantaneous capture of pneumonia in the problem list after CXR reports were available. Any significant delay in processing the reports would have resulted in reduced enrollment in the trial, or potential treatment decisions being made without the benefit of CDS. Additionally, utilizing the clinical trial teams and infrastructure allowed this implementation to benefit from extensive physician buy-in which can be difficult to achieve in other circumstances.

As mentioned above, we did not review all NLP+ encounters where CAP was added to the problem list from the BPA, nor all NLP- encounters where there was no evidence of pneumonia. We believe this was justified by the fact that, among those reviewed NLP+ encounters where CAP was not added to the problem list through the BPA, 95% had supporting radiology reports. Additionally, it would be unlikely for pneumonia not to have been added to the problem list or encounter diagnoses if the patient was diagnosed with pneumonia. Conversely, only 36% of reviewed NLP- encounters (those that had pneumonia problems or encounter diagnoses in the EHR) had reports that supported a pneumonia diagnosis. This was possibly due to time restriction of only scanning reports ordered in the first 24-hours after ED admission; these individuals may have had later CXR that supported diagnosis.

The most common errors in classification were false positives – classifying reports as supporting a diagnosis of pneumonia when they in fact did not. Only 16.5% of false positive encounters ever had some form of pneumonia added to the problem list; the majority of other false positive encounters were eventually diagnosed with different respiratory conditions (atelectasis, pleural effusion, bronchiolitis, etc.). Radiological findings for these diseases are, like pneumonia, often non-specific. These non-specific findings, along with possible (or "impossible to exclude") pneumonia were frequently mentioned in the findings section of the report, but the impression sections further specified a higher likelihood of other diagnoses. Sections of origin for UMLS concepts were not captured; had they been, it is possible the random forest models would have learned to weight the impression section higher. The lack of radiology report section information may have contributed to false positives in other ways, too. Reviewers noted that some reports had indication sections that mentioned suspected pneumonia, and that would have inappropriately been captured by the NLP as input to the random forest. Reviewers also noted that approximately 10% of false positive patients had findings of cystic fibrosis, sickle cell anemia, or acute chest syndrome. In retrospect, this is unsurprising since our training data was taken from a previous study,[2] and patients with these conditions were excluded from that study. Thus, our algorithm was not trained on radiology reports for patients with these conditions.

Inaccurate negation detection was also a likely driver of false positives. Reviewers noted that historical or resolving pneumonia was frequently mentioned in false positive reports, and likely was not captured correctly by the NLP. Phrases such as "interval resolution of the right apical opacity" were not recognized as negated. Phrases such as "infection not excluded," "underlying pneumonia difficult to definitively exclude", and "opacities favored to represent atelectasis over pneumonia," can be confusing for even human reviewers. They also noted that pneumonia was frequently mentioned in more specific terms (e.g., "right lower lobe pneumonia," "right middle lobe pneumonia," "multifocal pneumonia," bibasilar pneumonia, etc.), which would have been identified by the NLP as different UMLS concepts. Infrequently-used pneumonia concepts may not have been included in our model due to our methodology using the most-frequent UMLS concepts from the training set.

Fewer errors were due to false negatives. Out of 2433 NLP- encounters, only 2% were classified as false negatives after review. Of those, reviewers noted relevant terms such as "empyema" and "air bronchograms" that usually suggest possible pneumonia, but did not seem to have the desired impact on classification. As noted for false positive errors, this may be due to these terms being infrequent in our training set. Finally, a number of false negatives were in scoliosis patients where image-quality was deemed poor. It is unlikely that NLP could be used to compensate for poor image quality resulting in low-information CXR reports; however, additional training examples with limited information, appropriately categorized, could improve misclassification of similar reports.

Finally, the analysis revealed a number of areas where the algorithm could be improved in future work. Including section information in the random forest features would likely improve classification, as would additional training examples from patients with previously excluded conditions and low-quality imaging and reports. We also plan to enlarge the feature space to ensure all UMLS pneumonia concepts are included in future versions, as well as ensuring any radiological signs highly suggestive of pneumonia are included (in consultations with radiologists). The analysis of important features revealed a number of disparate pneumonia concepts, as well as similar concepts with distinct CUIs (e.g., multiple "opacity" concepts); we therefore plan to explore the use of concept normalization (i.e., binding similar concepts together, mapping rare pneumonia concepts to the base pneumonia concept) to improve recognition of pneumonia through the random forest. The most effective potential improvement, however, is likely in terms of negation detection. A thorough study of negation misclassification would help to identify additional trigger words to improve NegEx performance. Utilizing the ConText negation algorithm, instead of NegEx, may also improve negation and detection of historical conditions while only marginally increasing the processing time.[27,30] In addition to these NLP improvements, performing a thorough grid search to optimize the hyperparameters of the random forest would also likely improve prediction accuracy; other machine learning classifiers should be evaluated, as well. In the event of a future wider-scale implementation, we will also reconsider the 24-hours-after-ED-admission time restriction used in the clinical trial. Similarly, we used the earliest CXR reports for our training examples, as we wanted training examples to reflect early presentation of pneumonia; this will also need to be re-evaluated if the system is put into wider use. Finally, as pneumonia is not solely a radiological diagnosis, we would like to explore the addition of clinical symptoms to the model. As was shown in our analysis of important random forest features, *fever* and *cough* did contribute to model performance when mentioned in the CXR reports.

In conclusion, this work demonstrates that real-time NLP can effectively be integrated into the EHR to improve problem list completeness and delivery of CDS. While NLP and random forest models are effective at recognizing pediatric pneumonia from radiology reports, pneumonia is not solely a radiological diagnosis. The identification of supporting CXR reports, however, is an important component of most pediatric pneumonia diagnoses. By identifying such evidence, and triggering appropriate CDS when the diagnosis is confirmed by a provider, interventions such as this can expediate diagnosis and improve clinical care.

Acknowledgements

This research was supported by the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (R01AI125642). The authors gratefully acknowledge the ICE-CAP Study Team and members of the VUMC HealthIT team who contributed to this work, especially Leigh Price, Shari Just, and Dan Albert.

References

1. Dean NC, Jones BE, Jones JP, et al. Impact of an Electronic Clinical Decision Support Tool for Emergency Department Patients With Pneumonia. Ann Emerg Med. 2015 Nov;66(5):511–20.
2. Williams DJ, Zhu Y, Grijalva CG, et al. Predicting Severe Pneumonia Outcomes in Children. Pediatrics. 2016;138(4).
3. Improving CarE for Community Acquired Pneumonia (ICE-CAP) - ClinicalTrials.gov [Internet]. [cited 2020 Mar 24]. Available from: https://clinicaltrials.gov/ct2/show/NCT03760419
4. World Health Organization. Pneumonia Fact Sheet [Internet]. 2019 [cited 2020 Mar 13]. Available from: https://www.who.int/news-room/fact-sheets/detail/pneumonia
5. Katz SE, Williams DJ. Pediatric Community-Acquired Pneumonia in the United States: Changing Epidemiology, Diagnostic and Therapeutic Challenges, and Areas for Future Research. Infect Dis Clin North Am. 2018;32(1):47–63.
6. Leyenaar JK, Ralston SL, Shieh M-S, Pekow PS, Mangione-Smith R, Lindenauer PK. Epidemiology of pediatric hospitalizations at general hospitals and freestanding children's hospitals in the United States. J Hosp Med. 2016;11(11):743–9.
7. Jain S, Williams DJ, Arnold SR, Ampofo K, Bramley AM, Reed C, et al. Community-acquired pneumonia requiring hospitalization among U.S. children. N Engl J Med. 2015 Feb 26;372(9):835–45.
8. O'Grady K-AF, Torzillo PJ, Frawley K, Chang AB. The radiological diagnosis of pneumonia in children. Pneumonia Nathan Qld. 2014;5(Suppl 1):38–51.

9. Bradley JS, Byington CL, Shah SS, et al. The management of community-acquired pneumonia in infants and children older than 3 months of age: clinical practice guidelines by the Pediatric Infectious Diseases Society and Infectious Diseases Society of America. Clin Infect Dis Off Publ Infect Dis Soc Am. 2011;53(7):e25-76.
10. Novack V, Avnon LS, Smolyakov A, Barnea R, Jotkowitz A, Schlaeffer F. Disagreement in the interpretation of chest radiographs among specialists and clinical outcomes of patients hospitalized with suspected pneumonia. Eur J Intern Med. 2006 Jan;17(1):43–7.
11. Shah SN, Bachur RG, Simel DL, Neuman MI. Does This Child Have Pneumonia?: The Rational Clinical Examination Systematic Review. JAMA. 2017 Aug 1;318(5):462–71.
12. Haug P, Koehler S, Lau LM, Wang P, Rocha R, Huff S. A natural language understanding system combining syntactic and semantic techniques. Proc Symp Comput Appl Med Care. 1994;247–51.
13. Chapman WW, Haug PJ. Comparing expert systems for identifying chest x-ray reports that support pneumonia. Proc AMIA Symp. 1999;216–20.
14. Aronsky D, Haug PJ. An integrated decision support system for diagnosing and managing patients with community-acquired pneumonia. Proc AMIA Symp. 1999;197–201.
15. Aronsky D, Haug PJ. Automatic identification of patients eligible for a pneumonia guideline. Proc AMIA Symp. 2000;12–6.
16. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc JAMIA. 2000 Dec;7(6):593–604.
17. Aronsky D, Fiszman M, Chapman WW, Haug PJ. Combining decision support methodologies to diagnose pneumonia. Proc AMIA Symp. 2001;12–6.
18. Lagor C, Aronsky D, Fiszman M, Haug PJ. Automatic identification of patients eligible for a pneumonia guideline: comparing the diagnostic accuracy of two decision support models. Stud Health Technol Inform. 2001;84(Pt 1):493–7.
19. Liu V, Clark MP, Mendoza M, et al. Automated identification of pneumonia in chest radiograph reports in critically ill patients. BMC Med Inform Decis Mak. 2013 Aug 15;13:90.
20. Elkin PL, Froehling D, Wahner-Roedler D, et al. NLP-based identification of pneumonia cases from free-text radiological reports. AMIA Annu Symp Proc AMIA Symp. 2008 Nov 6;172–6.
21. Asatryan A, Benoit S, Ma H, English R, Elkin P, Tokars J. Detection of pneumonia using free-text radiology reports in the BioSense system. Int J Med Inf. 2011 Jan;80(1):67–73.
22. Makhnevich A, Sinvani L, Cohen SL, et al. The Clinical Utility of Chest Radiography for Identifying Pneumonia: Accounting for Diagnostic Uncertainty in Radiology Reports. AJR Am J Roentgenol. 2019;213(6):1207–12.
23. Guo W, Wang J, Sheng M, Zhou M, Fang L. Radiological findings in 210 paediatric patients with viral pneumonia: a retrospective case study. Br J Radiol. 2012 Oct;85(1018):1385–9.
24. Mendonça EA, Haas J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. J Biomed Inform. 2005 Aug;38(4):314–21.
25. Meystre S, Gouripeddi R, Tieder J, Simmons J, Srivastava R, Shah S. Enhancing Comparative Effectiveness Research With Automated Pediatric Pneumonia Detection in a Multi-Institutional Clinical Repository: A PHIS+ Pilot Study. J Med Internet Res. 2017 15;19(5):e162.
26. Breiman L. Random Forests. Mach Learn. 2001 Oct 1;45(1):5–32.
27. MetaMapLite [Internet]. [cited 2020 Mar 23]. Available from: https://metamap.nlm.nih.gov/MetaMapLite.shtml
28. MetaMap - A Tool For Recognizing UMLS Concepts in Text [Internet]. [cited 2019 May 10]. Available from: https://metamap.nlm.nih.gov/
29. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001 Oct;34(5):301–10.
30. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform. 2009 Oct;42(5):839–51.
31. The R Project for Statistical Computing [Internet]. [cited 2013 Apr 14]. Available from: http://www.r-project.org/
32. CRAN - Package randomForest [Internet]. [cited 2018 Jan 21]. Available from: https://cran.r-project.org/web/packages/randomForest/
33. Dinsdale et al. Supplemental Material [Internet]. [cited 2020 Jul 20]. Available from: https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.html