The effects of natural language processing on cross-institutional portability of influenza case detection for disease surveillance

Jeffrey P. Ferraro[1,2]; Ye Ye[3,4]; Per H. Gesteland[1,5]; Peter J. Haug[1,2]; Fuchiang (Rich) Tsui[3,4]; Gregory F. Cooper[3]; Rudy Van Bree[2]; Thomas Ginter[6]; Andrew J. Nowalk[7]; Michael Wagner[3,4]

[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA;
[2]Intermountain Healthcare, Salt Lake City, Utah, USA;
[3]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA;
[4]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, Pennsylvania, USA;
[5]Department of Pediatrics, University of Utah, Salt Lake City, Utah, USA;
[6]VA Salt Lake City Healthcare System, Salt Lake City, Utah;
[7]Department of Pediatrics, Children's Hospital of Pittsburgh of University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Keywords
Natural language processing, case detection, disease surveillance, generalizability, portability

Summary

Objectives: This study evaluates the accuracy and portability of a natural language processing (NLP) tool for extracting clinical findings of influenza from clinical notes across two large healthcare systems. Effectiveness is evaluated on how well NLP supports downstream influenza case-detection for disease surveillance.

Methods: We independently developed two NLP parsers, one at Intermountain Healthcare (IH) in Utah and the other at University of Pittsburgh Medical Center (UPMC) using local clinical notes from emergency department (ED) encounters of influenza. We measured NLP parser performance for the presence and absence of 70 clinical findings indicative of influenza. We then developed Bayesian network models from NLP processed reports and tested their ability to discriminate among cases of (1) influenza, (2) non-influenza influenza-like illness (NI-ILI), and (3) 'other' diagnosis.

Results: On Intermountain Healthcare reports, recall and precision of the IH NLP parser were 0.71 and 0.75, respectively, and UPMC NLP parser, 0.67 and 0.79. On University of Pittsburgh Medical Center reports, recall and precision of the UPMC NLP parser were 0.73 and 0.80, respectively, and IH NLP parser, 0.53 and 0.80. Bayesian case-detection performance measured by AUROC for influenza versus non-influenza on Intermountain Healthcare cases was 0.93 (using IH NLP parser) and 0.93 (using UPMC NLP parser). Case-detection on University of Pittsburgh Medical Center cases was 0.95 (using UPMC NLP parser) and 0.83 (using IH NLP parser). For influenza versus NI-ILI on Intermountain Healthcare cases performance was 0.70 (using IH NLP parser) and 0.76 (using UPMC NLP parser). On University of Pisstburgh Medical Center cases, 0.76 (using UPMC NLP parser) and 0.65 (using IH NLP parser).

Conclusion: In all but one instance (influenza versus NI-ILI using IH cases), local parsers were more effective at supporting case-detection although performances of non-local parsers were reasonable.

Correspondence to
Jeffrey P. Ferraro
Homer Warner Center | Intermountain Healthcare
5171 South Cottonwood St, Suite 220
Murray, Utah 84107
Jeffrey.Ferraro@imail.org
Tel: 801–244–6570

Appl Clin Inform 2017; 8: 560–580
https://doi.org/10.4338/ACI-2016-12-RA-0211
received: December 31, 2016
accepted: March 11, 2017
published: May 31, 2017
Citation: Ferraro JP, Ye Y, Gesteland PH, Haug PJ, Tsui F(R), Cooper GF, Van Bree R, Ginter T, Nowalk AJ, Wagner M. The effects of natural language processing on cross-institutional portability of influenza case detection for disease surveillance. Appl Clin Inform 2017; 8: 560–580
https://doi.org/10.4338/ACI-2016-12-RA-0211
Funding
Research reported in this publication was supported by grant R01LM011370 from the National Library of Medicine.

1. Objective

Surveillance of a population for infectious and other diseases is an important public health function. Public health has historically relied on clinical diagnosis, i.e., educating clinicians to report individuals with certain diseases when they are diagnosed. However, due to incompleteness and time delays in clinician-based reporting, there has been substantial interest in accomplishing the same task using automated algorithms that analyze information collected by electronic medical records [1].

A key technical barrier to automatic case detection from electronic medical records has been the lack of coded information about symptoms and signs of disease, which clinicians at present record in unstructured free-text clinical notes. Natural language processing (NLP) can be put to use on this problem, turning the unstructured information contained within clinical notes into machine-readable, coded data.

An open problem with using NLP is that clinical notes take many forms and differ in content and structure across institutions. Even within a single institution these characteristics may change over time causing system performance to drift. For an NLP-based approach to be useful in population disease surveillance its portability characteristics must be understood. To some degree, NLP tools must be resilient to local differences if they are to be effective in providing information to support widescale disease surveillance.

This study evaluates the accuracy and portability of a natural language processing tool for extracting clinical findings of influenza across two large healthcare systems located in different regions of the United States. Two locally developed NLP parsers, one at each institution were developed using local institution clinical notes from emergency department (ED) encounters indicating influenza, and then evaluated at both institutions. The effectiveness of each parser is evaluated on how well it supports downstream case detection of influenza which is a critical component of disease surveillance for public health outbreak detection.

2. Background and Significance

Disease surveillance and outbreak detection are fundamental activities to assist in early public health management and response to bioterrorism threats and infectious disease outbreaks like influenza [2–5]. New strains of old diseases and new diseases continue to emerge that require ongoing public health vigilance [6, 7]. Early outbreak detection systems that can be quickly deployed on a widescale can help address the need for rapid response providing effective outbreak management [8]. Limited attempts and successes have been realized in portability of outbreak detection systems [2, 9, 10]. The Real-time Outbreak and Disease Surveillance system (RODS) developed at University of Pittsburgh was successfully implemented in Salt Lake City, Utah during the 2000 Olympics for surveillance of biological agents such as Anthrax [11]. At the national level, the BioSense platform [12] and the Essence system [13] have incorporated some standardized tooling that can collect, evaluate, and share syndromic surveillance information among health officials and government agencies. To our knowledge, these are the few automated biosurveillance systems that have tried to address system interoperability and standardization across institutional boundaries.

An important new input to disease surveillance systems is the information contained within unstructured clinical notes at treating healthcare institutions. At present, disease reporting systems depend on clinicians to establish and report patients with selected diagnosis. Recent work has shown that diagnosis can be inferred from clinical notes [14–20]. Natural language processing (NLP) is the essential component used to extract relevant clinical signs and symptoms from unstructured clinical notes that help to identify the presenting syndrome in real-time surveillance systems [21, 22].

Clinical natural language processing techniques used for extraction of clinical findings from clinical notes have advanced over the last decade to make NLP a viable operational component in clinical systems [14, 23–25] and Biosurveillance systems [26]. Unfortunaly, these systems are usually developed in localized settings addressing the needs of single healthcare institutions. Clinical notes vary to such an extent across healthcare systems that NLP components are typically over-fit to the target institution and do not generalize well when migrated to other healthcare settings [27, 28]. This issue limits the ability to share NLP components across institutional boundaries. Research and

method advancements in areas like domain adaptation hold promise in addressing this limitation in the hope that these systems can one day be shared and rapidly deployed across institutional boundaries [29–31].

On the other hand, we need to better understand the performance characteristics of clinical NLP components on downstream processes. NLP components are expensive to develop and even more costly to develop in a generalized way. To our knowledge, very few studies have been done to determine how good is good enough when it comes to generalizability of NLP components and their impact on downstream processes in clinical pipeline systems [32, 33]. This study examines this issue where extraction of clinical findings using NLP is the source information supporting downstream case detection of influenza for outbreak detection and disease surveillance.

To conduct this study we used an NLP tool developed at University of Pittsburgh called Topaz [34]. Topaz is a pipeline system that extracts domain specific clinical findings and their modifiers from clinical notes using deduction-rules. Modifiers include whether the finding is absent or present, recent or historical, and whether experienced by a patient or someone else such as a family member. Deterministic production rules are constructed based on clinical text patterns that suggest the clinical finding of interest. The pattern-matching is expressed in the form of regular expressions which describe the text search patterns. These search expressions act as preconditions that fire actions forming a production rule. The system supports forward-chaining of production-rules so that complex expressions and inferences can be made on clinical text [35].

Topaz supports conflict resolution when a finding is identified as both absent and present in different segments of the same clinical note. It resolves the conflicting finding assertions in favor of being present when both absent and present assertions have been made. Topaz has four main modules, (1) a structural document preprocessor to identify clinical note sections, sentence boundaries, and tokenization, (2) a UMLS Metathesaurus [36] concept mapper with extension capabilities, (3) a forward-chaining deductive inference engine, and (4) a conflicting-concept resolver. Topaz has been evaluated in several studies that have reported reasonable operating characteristics [17, 34, 37, 38].

Other successful applications of rule-based NLP approaches to information extraction have surfaced recently that use similar extraction techniques to that of Topaz [39, 40]. MedTagger [41] follows a similar processing paradigm to Topaz but it is integrated into the clinical Text Analysis and Knowledge Extraction System (cTAKES) [42] as a component of this framework. Topaz is a complete pipeline system that is similarly built using the Unstructured Information Management Architecture (UMIA) framework [43] as does cTAKES, although its primary focus is information extraction. MedTagger is an information extraction component that relies on several of the other cTAKES components for processing tasks like sentence splitting, section identification, and negation. Both pipelines involve similar processing tasks although they package their processing components differently. Topaz comes with default section header rules of its own which can be extended for adaptation to institutional variances while MedTagger relies on SecTag [44] for section header identification. Both systems can activate or deactivate the use of section header identification although in Topaz this is done by directly inactivating certain rule definition files rather than reconfiguring the pipeline itself. Although both tools use regular expressions for matching, Topaz integrates the regular expression matching into production rules that are executed by a more advanced forward chaining deduction engine. This allows Topaz the ability to perform very complex rule-chaining, supporting deeper inference capabilities in the information extraction process.

In this study, we used a Bayesian case detection system (CDS) developed at University of Pittsburgh to classify each ED encounter as influenza, non-influenza influenza-like illness (NI-ILI), and 'other' based on symptoms and findings extracted by Topaz. Integral to this system are Bayesian network [45] models and an inference engine that results in disease classification [17, 34, 38]. CDS takes as input, the clinical findings (F) for a patient case that are produced by Topaz. It outputs the posterior probability distribution that the patient has one of several diseases given those findings. The diseases D represented as a Bayesian network model in CDS are influenza, NI-ILI, and 'other' which is represented as one broad category. CDS performs probabilistic case detection by using a Bayesian network model to compute $P(D|F)$, the posterior probability of the disease given the findings. A Bayesian network is a graphical model representation which provides a method of reasoning under uncertainty. The nodes of a Bayesian network represent variables and arcs between the nodes represent conditional dependencies between these variables. The strength of the relationships be-

tween the nodes (variables) are represented as conditional probability distributions. Bayesian networks models factorize a joint probability distribution as the product of its conditional probability distributions, which often yields a compact representation of the joint distribution. Bayesian networks have shown to be well-suited for clinical diagnostic prediction where only a portion of the target clinical features may be available on a patient case as they are very robust to missing data [15, 17, 46].

3. Materials and Methods

Institutional Review Board (IRB) approval was obtained from both healthcare systems prior to conducting this study.

3.1 Healthcare Systems Characteristics

Intermountain Healthcare (IH) is the leading integrated health care delivery system in Utah. The health system operates 22 community, tertiary, and specialty hospitals, a health plan, and 1,400 employed physicians. Intermountain also operates 185 clinics, including primary care clinics, and urgent care clinics. Intermountain has 137,000 annual admissions and 502,000 annual emergency department (ED) visits across the entire system.

The University of Pittsburgh Medical Center (UPMC) is closely affiliated with its academic partner, the University of Pittsburgh and is the leading integrated health care delivery system in western Pennsylvania. UPMC operates 25 academic, community, tertiary, and specialty hospitals, a health plan, and 2,500 employed physicians. UPMC has 170,000 annual admissions with estimated 58% market share for Allegheny County and 720,000 annual emergency department visits.

3.2 NLP Study Datasets

We constructed four datasets to carry out the NLP parser experiments in this study. From each Healthcare System (IH and UPMC), one development/training corpus and one test corpus were constructed.

3.2.1 Intermountain Healthcare NLP Datasets

The Intermountain Healthcare datasets were selected from ED encounters across 19 of their facilities. The inclusion criteria consisted of positive influenza cases by microbiology culture, direct fluorescent-antibody (DFA) testing, or polymerase chain reaction (PCR) testing with at least one ED Physician or Licensed Independent Practitioner report. The first report was selected for encounters with multiple physician notes.

The development/training corpus was constructed from the first 100 adult (age >= 6) influenza cases spanning January 1st, 2007 – February 28th, 2008 and 100 pediatric (age < 6) cases spanning January 2nd, 2007 – February 16th, 2007, totaling 200 distinct cases. The test corpus was constructed using the next consecutive clinical encounters. This corpus was made up of 100 adult influenza cases spanning March 2nd, 2008 – June 8th 2009 and 100 pediatric cases spanning February 17th, 2007 – March 20th, 2007, totaling 200 distinct cases.

3.2.2 University of Pittsburgh Medical Center NLP Datasets

The University of Pittsburgh Medical Center datasets were selected from ED encounters across 5 EDs: UPMC Presbyterian Hospital, UPMC Shadyside Hospital, UPMC McKeesport Hospital, UPMC Mercy Hospital, and Children's Hospital of Pittsburgh of UPMC. The inclusion criteria consisted of positive influenza cases confirmed by polymerase chain reaction (PCR) testing with at least one clinician report. The earliest signed clinical report was selected for encounters with multiple clinician reports.

The development/training corpus was constructed from the first 100 adult (age >= 6) influenza cases spanning March 15th, 2007 – February 27th, 2008 and 100 pediatric (age < 6) cases spanning December 21st, 2007 – October 20th, 2009, totaling 200 distinct cases. The test corpus was con-

structed using the next consecutive clinical encounters. This corpus was made up of 100 adult influenza cases spanning February 28th, 2008 –March 26th, 2009 and 100 pediatric cases spanning October 20th, 2009 – February 12th, 2011, totaling 200 distinct cases.

3.3 Annotation Process

Three board-certified practicing physicians (one internist, and two pediatricians from each institution) identified 77 clinical findings by process of consensus covering the four diseases of study specified in the research design – influenza, respiratory syncytial virus (RSV), metapneumovirus, and parainfluenza. They identified the clinical findings based on experience treating cases of influenza within their respective institutions. Seventy of these clinical findings were relevant to influenza as shown in ▶ Appendix A. The other three diseases would be studied in later research. Annotation of a clinical finding involved identifying the clinical finding as either absent or present in the clinical note and marking the text phrase indicating the finding. An outside, independent annotation service (University of Utah Core Research Lab) was contracted to provide annotation services for this study. Four licensed RNs were trained as annotators from a master annotation guideline providing the clinical finding definition accompanied with example phrases, utterances, and lexical variants commonly documented by treating clinicians for each clinical finding. The eHOST (Extensible Human Oracle Suite of Tools) [47] open source annotation tool was used for annotation. Training was performed using 80 (40 adult/40 pediatric) reports randomly selected from the training corpus of each of the two healthcare systems, IH and UPMC. This represented a total of 160 annotated training cases, 80 from each site. The annotation training was conducted over four rounds, each consisting of 20 clinical notes in each round. For each round, two randomly selected annotators were given the same set of 20 reports and kappa was calculated between those annotators to assess consistency. Discrepancies between annotator pairs were adjudicated by the physician board-certified in internal medicine and feedback was provided. The four rounds of annotation training resulted in kappa scores above 0.80 between annotator pairs.

Focus was then turned to the test corpora. The test corpora consisted of 200 (100 adult/100 pediatric) clinical notes from each healthcare system broken into 5 paired annotation sets containing 24 reports per individual set. Each paired set had 8 duplicate reports contained within the pair so that inter-annotator agreement across paired annotators could be measured. So for any paired annotation set, this represented 40 distinct reports – 16 unique reports in each individual annotation set, and an additional 8 duplicated reports across the pair. The annotators were assigned annotation sets such that all four annotators would be equally paired with one another across all of the annotation sets. This process resulted in 200 (100 adult/100 pediatric) distinct annotated reports being generated for each healthcare system for testing purposes. Inter-annotator agreement using Fleiss' kappa [48] across the 80 shared reports representing 20% of the reports was measured and reported. These 80 duplicated clinical notes across annotation sets for purposes of measuring kappa were then adjudicated by a board-certified physician.

3.4 NLP Parser Development

Topaz [34] is a rule-based natural language processing parser capable of extracting clinical concepts by applying domain specific pattern-matching and deduction rules. Two influenza parsers were developed separately by two different development teams blinded to one another's development activities. One team was provided the IH training corpus while the other team was given the UPMC training corpus. Each training corpus contained the 80 annotated clinical notes and the additional 120 unannotated notes. Each team consisted of an experienced NLP software engineer and a board certified pediatrician on staff from each institution. No communications between teams was permitted. They were allowed to evaluate their respective parsers on the local annotated training corpus provided to each team as often as deemed necessary but were unable to evaluate their parser against each other's corpus during the development phase of the study. Once each team felt that their parser's operating characteristics were optimal, the systems were evaluated one time against the test corpus of each healthcare system to determine the cross-compatibility performance characteristics.

Recall, precision, and F_1-score were measured for local healthcare system performance characteristics and cross-site compatibility characteristics.

3.5 CDS Study Datasets

At the core of CDS are Bayesian network classifier models that perform case detection based on the clinical findings extracted from the clinical notes by the NLP parsers. We built four Bayesian network classifiers that differed in source training data (IH or UPMC clinical notes) and NLP parser (IH NLP parser or UPMC NLP parser) to extract the clinical findings. The training datasets were used to develop the Bayesian network models using machine learning to learn the network structure as well as the joint probability distributions of the models. We used the K2 algorithm [49] to machine learn the structure of the models from this training data. The K2 algorithm uses a forward-stepping, greedy search strategy to identify the conditional dependency relationships (arcs) among the variables (nodes) of a Bayesian network that produce a locally optimal network structure [49].

We labeled as *influenza,* patient encounters with a positive laboratory test for influenza by polymerase chain reaction (PCR), direct fluorescent antibody (DFA), or viral culture. Among the remaining encounters, we labeled as *NI-ILI* cases with at least one negative test for PCR, DFA, or culture. All remaining encounters were labeled as *other.*

3.5.1 Intermountain Healthcare CDS Training Datasets

The Intermountain Healthcare CDS training dataset consisted of 47,504 ED encounters between January 1, 2008 and May 31, 2010, including 1,858 *influenza*, 15,989 *NI-ILI, and* 29,657 *other* encounters. The IH training dataset represented 60,344 clinical notes. When an encounter was associated with more than one clinical note, we used the union of clinical findings from all of the clinical notes.

3.5.2 University of Pittsburgh Medical Center CDS Training Datasets

The University of Pittsburgh Medical Center CDS training dataset consisted of 41,189 ED encounters drawn from the same period as Intermountain Healthcare and labeled using the same criteria. This training dataset included 915 *influenza*, 3,040 *NI-ILI, and* 37,234 *other* encounters. The UPMC training dataset was associated with 76.467 clinical notes. Again, for encounters with multiple notes we would union the clinical findings from all of the clinical notes.

3.5.3 Healthcare Systems' CDS Test Datasets and Evaluation

To test downstream CDS performance based on the signs and symptoms extracted using the healthcare systems' NLP parsers (IH NLP Parser and UPMC NLP Parser), we collected one year of ED visits as a test dataset from each healthcare system spanning June 1st, 2010 – May 31st, 2011. There were 220,276 IH ED clinical notes representing 182,386 ED visits and 480,067 UPMC ED notes representing 238,722 ED visits. The ED reports collected from each healthcare system were parsed by the IH NLP parser and the UPMC NLP parser producing four data sets, (1) IH reports parsed by IH NLP parser, (2) IH reports parsed by UPMC NLP parser, (3) UMPC reports parsed by UPMC NLP parser, and (4) UPMC reports parsed by the IH NLP parser. The CDS Bayesian networks were then evaluated using these datasets to determine the effects of NLP parser cross-compatibility on downstream case detection.

4. Results

4.1 Inter-annotator Agreement on NLP Datasets

Inter-annotator agreement was measured using Fleiss' kappa [48]. On the Intermountain Healthcare test data set an agreement score of 0.81 (95% CI: 0.80 to 0.81) was achieved over 1,477 clinical findings. On the University of Pittsburgh Medical Center test data set a kappa score of 0.81 (95% CI: 0.80 to 0.82) over 1,504 clinical findings was achieved. For both data sets reliable agreement was reached.

4.2 Healthcare System Differences in Language Expressing Clinical Findings

To gain insight into the differences in the language characteristics used by each healthcare system to describe a particular clinical finding, the annotated text from the test corpora that describes a clinical finding as absent or present in a clinical note was analyzed to identify statistically significant distributional differences that may exist across healthcare systems. For example, the clinical finding *chest wall retractions – present*, one institution may describe in a note as "*was using his accessory muscles for respiration*" while the other institution may describe it as "*does have some abdominal breathing and moderate subcostal and mild to moderate suprasternal retractions*". ▶Figure 1 describes the results of Fisher's Exact Test for homogeneity performed on the word frequencies describing each clinical finding between healthcare systems. For each healthcare system, the annotated text segments indicative of a clinical finding being absent or present found in the clinical notes were broken down into a bag of words with frequency counts. This was done for each clinical finding. The test statistic was then applied between the bag of words frequency counts representing each clinical finding between the institutions. This evaluation was performed on the annotated test corpora with word stemming, and stop words removed. P-value significance was adjusted for multiple comparisons testing using the false discovery rate [50] method. Seventy percent (n = 49/70) of the clinical findings had statistically significant (adjusted p-value < 0.05) differences in the language used to express the clinical findings by each healthcare system.

4.3 Clinical findings distributional characteristics between institutions

The frequency distribution of annotated clinical findings indicating influenza across the test corpora of both institutions is illustrated in ▶Figure 2. ▶Appendix A also provides the frequency counts of the annotated clinical findings. Seventy clinical findings were identified by expert annotation across the test corpus for each institution. Evaluation of the distributional characteristics was analyzed by calculating a multinomial chi-square goodness of fit [51] test statistic. The test statistic produced a p-value < 0.0001 confirming that the two frequency distributions are drawn from different distribution functions. This finding may imply that (1) the signs and symptoms documented by clinicians in the clinical notes to describe influenza cases differ between institutions or (2) there may be differing influenza patient presentations between institutions. The top ten most frequent clinical findings found in the clinical notes for each institution is shown in ▶Table 1.

4.4 NLP parser performance for influenza clinical findings

▶Table 2 presents the performance characteristics for each NLP parser using the rule set that was developed from the institution's development/training corpus but evaluated on the test corpora from both institutions. In other words, the IH NLP parser extraction rules were constructed using information contained within the Intermountain Healthcare training corpus and the UPMC NLP parser extraction rules were constructed using information contained within the University of Pittsburgh Medical Center training corpus. The goal of this experiment was to evaluate the performance loss in applying locally developed extraction rule sets against foreign (non-local) clinical notes to assess generalizability and determine the downstream effects of any upstream performance loss. On the Intermountain Healthcare test corpus, recall, precision, and F_1-score of the IH NLP parser were 0.71 (95% CI: 0.70 to 0.72), 0.75 (95% CI: 0.73 to 0.76), and 0.73 respectively, and the UPMC NLP parser, 0.67 (95% CI: 0.65 to 0.68), 0.79 (95% CI: 0.78 to 0.80), and 0.73. For the Intermountain Healthcare corpus, the local IH NLP parser had statistically significant (p-value < 0.0001) better recall than the non-local UPMC NLP parser, but statistically significant (p-value < 0.0001) lower precision. The F_1-scores were equal. It is difficult to determine which of the parsers may be better suited in this instance as it would depend on whether precision or recall is more important depending on the target disease differentiation. On the University of Pittsburgh Medical Center test corpus, recall, precision, and F_1-score of the UPMC NLP parser were 0.73 (95% CI: 0.71 to 0.74), 0.80 (95% CI: 0.79 to 0.82), and 0.76 respectively, and the IH NLP parser, 0.53 (95% CI: 0.51 to 0.54), 0.80 (95% CI: 0.78 to 0.81), and 0.64. For the UPMC corpus, the local UPMC NLP parser had statistically significant (p-value <

0.0001) better recall than the non-local IH NLP parser and equivalent precision although the F_1-score of the UPMC NLP parser was quite a bit better. In this case, the local UMPC NLP parser with higher recall and equivalent precision would suggest operational preference over the non-local IH NLP parser. In both compatibility comparisons the local parser may be preferred over the non-local parser.

4.5 Downstream case detection performance with NLP parsers

The effects of the NLP parsers on downstream case detection were evaluated based on the ability of the Bayesian case detection system to discriminate between influenza and non-influenza, as well as, the more difficult case of influenza and non-influenza influenza-like illness. ▶Figure 3 illustrates the effects of NLP parser performance on case detection by comparing area under the receiver operating characteristic curves (AUROCs). Significance was calculated using DeLong's [52] statistical AUROC comparison test. The AUROCs and comparison test statistic results are shown in ▶Table 3. For discriminating between influenza and non-influenza, the two parsers supported case detection almost equivalently on Intermountain Healthcare cases with an AUROC of 0.932 for the IH NLP parser and 0.936 for the UPMC NLP parser. In detecting influenza from non-influenza on University of Pittsburgh Medical Center cases, the local UPMC NLP parser (AUROC = 0.954) outperformed the IH NLP parser (AUROC = 0.843). For the more difficult discrimination of influenza from NI-ILI on IH cases, the non-local UPMC NLP parser (AUROC = 0.748) outperformed the local IH NLP parser (AUROC = 0.698). This may be due to precision being more important than recall for distinguishing between cases of influenza and NI-ILI. This seems to make intuitive sense in that influenza and NI-ILI have more similarities in their disease presentations than influenza versus non-influenza making precision more important in distinguishing among the cases. On the University of Pittsburgh Medical Center cases, the local UPMC NLP parser better supported discrimination between influenza and NI-ILI with an AUROC of 0.766. In all but one instance (influenza versus NI-ILI using IH cases), the local parsers seemed to do better or as good a job of supporting downstream case detection then that of the non-local parsers. Although the results produced by the non-local parsers may still be considered within reasonable limits if there is a need for rapid and widespread surveillance deployment.

5. Discussion

As this study illustrates, it is common to find language differences in clinical notes describing clinical findings among institutions. Even within institutions, dictation styles and linguistic expression may vary among clinicians. These variations are typically addressed in locally developed NLP systems because a representative sample of the variation can be obtained. Yet this does not typically address the increased variation experienced across institutional boundaries. This is one of the most difficult challenges faced in generalizing modern NLP systems today. Whether rule-based or statistically based, NLP systems are developed from training sets providing samples of phrases and linguistic expression to draw upon in developing extraction rules or statistical extraction methods.

A significant difference in NLP rule-development between Intermountain Healthcare and University of Pittsburgh Medical Center was that the IH rule-developer considered semi-structured section headers in an effort to gain an early context before applying extraction rules. These section header identification rules addressed roughly 150 section header variants within the IH dataset alone. Post study analysis determined that the UPMC dataset had 15 section header types, none of which were syntactically or semantically consistent enough with Intermountain section headers to be identified by the IH NLP parser. Section header identification relies heavily on syntactical attributes like capitalization, ending punctuation, number of line feeds, and structural aspects like section order, in addition to the lexical content [44]. This attributed to the decreased performance of the IH NLP parser when ran against the UPMC clinical notes. Anticipated clinical note sections identifiable by these rules were not recognized within the UPMC notes and therefore large segments of clinically relevant text were ignored by the IH NLP parser.

Surprisingly, the UPMC NLP parser with rules that did not take this approach produced better precision then the IH NLP parser but worse recall when ran against the IH dataset. This led to the UPMC NLP parser doing a better job at supporting downstream case detection than the IH NLP parser. In the case of the UPMC NLP parser, all of the text segments within the clinical notes were processed for relevant clinical findings, regardless of institution. Besides this difference in approach, both rule sets used regular expression matching with negation, deductive inference to identify the absence or presence of a clinical finding, and the default conflict resolution logic when a finding is identified as both absent and present in different segments of the same clinical note. The default conflict resolution is to favor an assertion of present over absent.

A deeper analysis of the clinical finding extraction rules revealed that the IH NLP parser rules were more specific in nature than the UPMC NLP parser rules. The IH rules included surrounding context words while the UPMC NLP parser rules were in many instances simple clinical finding terms without surrounding context. The effects of rule specificity related to performance would be consistent with our findings of statistically significant differences in language expressing the clinical findings between the two institutions. By considering less surrounding context there is less of a chance that the language differences would affect performance as seen with the UPMC NLP parser.

Systems that show reasonable generalizing characteristics have typically done a good job at incorporating guessing heuristics into their extraction algorithms. These guessing heuristics are used in anticipation of coming across phrase expressions that were not seen in the development/training corpus. In information extraction, it is challenging to develop good guessing heuristics because there is typically over-fitting caused by the narrow lexical scope of examined phrases from the local development corpus to express a clinical concept and the extraction rules. It is also very difficult to develop general extraction methods to address synonymy at a complex phrase level. As identified between Intermountain Healthcare and University of Pittsburgh Medical Center, generalizability is further complicated in that patient disease presentations or how clinicians express clinical findings among patient cases may differ.

Another approach to better addressing generalizability may be the use of terminology and ontology mapping tools such as MetaMap [53] to map clinical text to standard UMLS Metathesaurus [36] concepts that could be used to identify common findings in clinical notes across institutions. This may help to provide an interoperability bridge improving NLP generalizability. One of the limitations with this approach is that text from clinical notes have been shown to be their own sublanguage [54]. In a continuing study on domain adaptation of part-of-speech tagging for clinical notes that took advantage of lexical content, the SPECIALIST lexicon only provided 48.7% vocabulary coverage across a clinical note corpus made up of the ten most common clinical note types at Intermountain Healthcare [55]. The SPECIALIST lexicon is one of the foundational components of MetaMap. More work needs to go into developing terminology and ontologies that have coverage for the clinical sublanguages found in clinical notes.

The encouraging message is that even when there is limited opportunity for generalizing certain natural language processing tasks our findings suggest that downstream processes may still operate within reasonable limits. Although we did experience a drop off in NLP performance when applying some of the locally developed parsers across institutional boundaries, the downstream task of case detection was still able to produce good results in differentiating cases of influenza and non-influenza. The more difficult task of identifying influenza from non-influenza influenza-like illness resulted in lower performance, but surprisingly, the UPMC NLP parser (non-local parser) better supported this case detection task when applied to Intermountain Healthcare cases. This may be due to the UPMC NLP parser having better precision than the IH NLP parser on IH clinical notes. This performance characteristic may be more important to distinguish diseases with very similar signs and symptoms. This finding is encouraging and supports our optimism that systems can be ported across institutional boundaries with reasonable operating characteristics.

We developed the CDS system in large part to provide probabilistic information about each patient case to an Outbreak Detection and characterization Systems (ODS) that we developed [56]. ODS uses that information to detect and characterize disease outbreaks. The information is a likelihood of the form *P(evidence | disease)*, where "evidence" is a set of clinical finding variables, and "disease" can be one of a number of disease states, including for example influenza. A Bayesian network provides a natural way to represent evidence and diseases and to infer the needed likelihoods.

Most other machine learning methods, such as random forests, do not provide a direct way to derive such likelihoods, rather, those methods are intended to directly derive posterior probabilities of the form *P(disease | evidence)*. We have also performed previous studies evaluating Bayesian networks against other top machine learning algorithms like random forests that have shown that Bayesian networks perform comparable to these other top learning methods [57].

5.1 Limitations

The natural language processing tool used in this study takes a rule-based approach to information extraction and was not well suited for experimenting with more advanced statistically based NLP information extraction techniques. Some statistical approaches open the door to exploring methods of unsupervised or semi-supervised domain adaptation where source models can self-adapt to distributional characteristics of new domains [58–60]. Due to limited available syndromic disease data across these two healthcare systems and limited resources, we were unable to expand our compatibility research beyond that of influenza. Also, compatibility studies involving more institutions need to be performed to further assess the issues of interoperability and to draw further insights into the challenges faced in porting natural language processing and Biosurveillance systems across institutional and geographical boundaries.

6. Conclusion

Portability and rapid deployment of infectious disease surveillance systems across geographic and institutional boundaries require tools that are resilient to local knowledge representation differences for effective surveillance to take place. This research addresses the important question of cross-institutional portability of natural language processing systems to support disease surveillance. Our results suggest that there is still the need for further research in methods development to produce more generalizable NLP tools. Natural language processing is becoming an integral disruptive technology to support surveillance systems although this study suggests it is sensitive to institutional variability. To our knowledge, this is one of the few comprehensive studies done on portability across institutional boundaries. There is a compelling need for more of these sorts of study designs to be carried out in sub-systems like clinical natural language processing which can lead to more robust and effective Public Health system solutions. Our research concludes that portability of systems that incorporate NLP can be achieved to some degree but more work needs to be done in this area.

Clinical Relevance

Disease surveillance and outbreak detection are critical activities to assist in early public health management and response to bioterrorism threats and infectious disease outbreaks [2–5]. Portability and rapid deployment of infectious disease surveillance systems across geographic and institutional boundaries require tools that are resilient to local knowledge representation differences for effective surveillance to take place. This research addresses the important question of cross-institutional portability of natural language processing systems to support disease surveillance.

Questions

1. Generalization of natural language processing (NLP) tools for use across institutional boundaries can be challenging because?

A) There is greater variability in dictation styles and the linguistic expression found in clinical notes across institutional boundaries than within single institutions.

B) NLP tools are better off being over-fit to local institutional clinical note variances because this improves the operational performance of these tools.

C) Domain adaptation of NLP tools have not shown much promise in providing an avenue for improved generalization.
D) The common use of terminology services to support natural language processing make generalization difficult.

Answer: A)

As this study illustrates, it is common to find language differences in clinical notes describing clinical findings among institutions. Even within institutions, dictation styles and linguistic expression may vary among clinicians. These variations are typically addressed in locally developed NLP systems because a representative sample of the variation can be obtained. Yet this does not typically address the increased variation experienced across institutional boundaries. This is one of the most difficult challenges faced in generalizing modern NLP systems today. Whether rule-based or statistically based, NLP systems are developed from training sets providing samples of phrases and linguistic expression to draw upon in developing extraction rules or statistical extraction methods.

2. Disease surveillance and outbreak detection are important public health functions because?
A) Surveillance systems can help to provide the necessary information to public health officials for more effective outbreak management.
B) New infectious and non-infectious pathogens continue to emerge that require ongoing public health awareness.
C) For the improved national and public safety against bioterrorism.
D) All of the above

Answer: D)

Disease surveillance and outbreak detection are fundamental activities to assist in early public health management and response to bioterrorism threats and infectious disease outbreaks like influenza. New strains of old diseases and new diseases continue to emerge that require ongoing public health vigilance. Early outbreak detection systems that can be quickly deployed on a widescale can help address the need for rapid response providing effective outbreak management.

Conflicts of Interest

The authors declare that they have no conflicts of interest relevant to this research.

Human Subjects Protection

This study was conducted with Institutional Review Board (IRB) approval obtained from both healthcare systems governing protection of human and animal subjects.

Acknowledgements

The authors would like to thank Dr. Wendy Chapman for her consultation on the Topaz NLP parser framework. We would also like to thank Lee Christensen for his work operationalizing the Topaz NLP parser for this study.

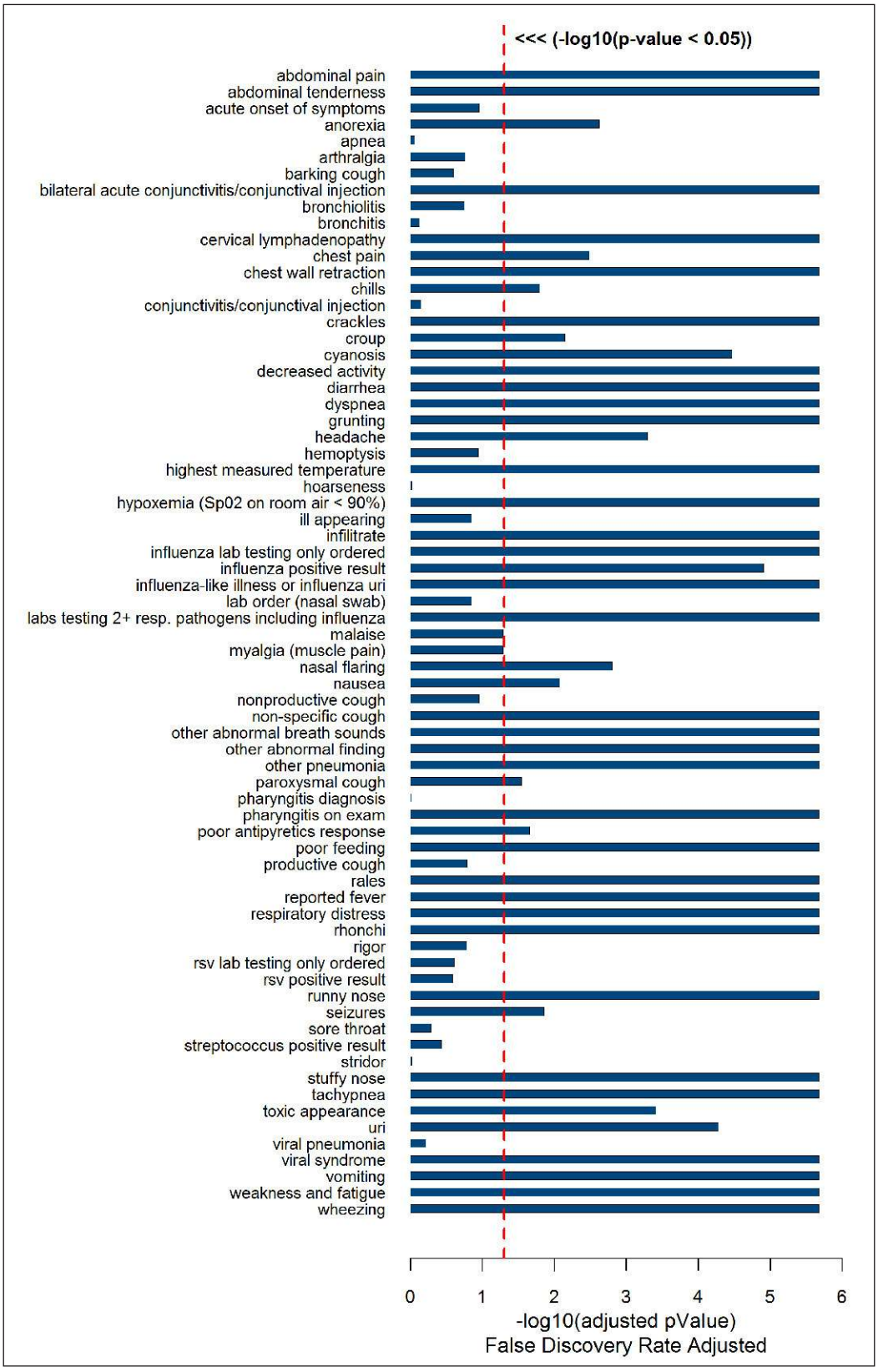

Fig. 1 Language differences expressing clinical findings for Influenza between Intermountain Healthcare and University of Pittsburgh Medical Center. Adjusted significance is shown on a logarithmic scale.

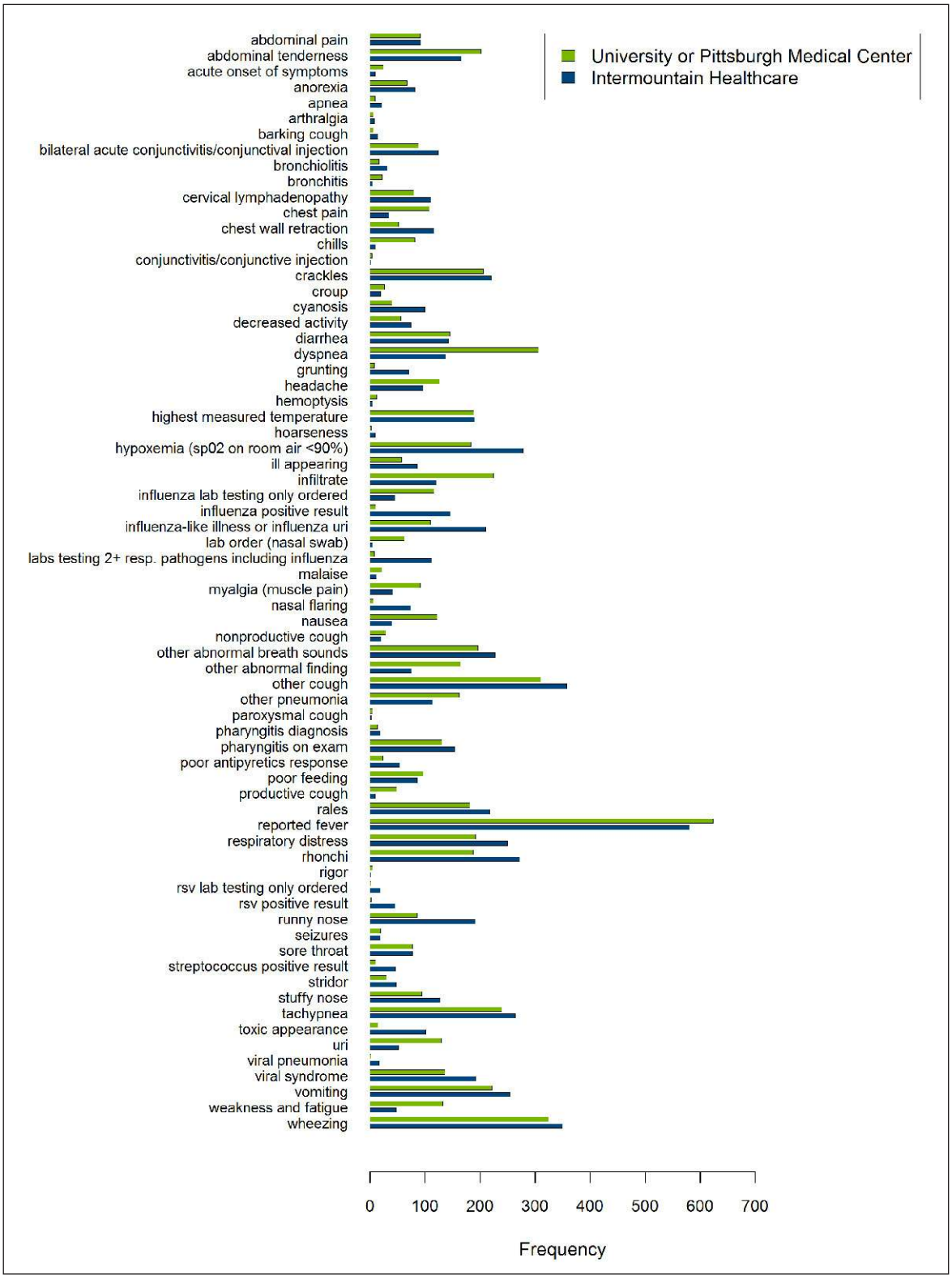

Fig. 2 Frequency distribution of annotated clinical findings for Influenza.

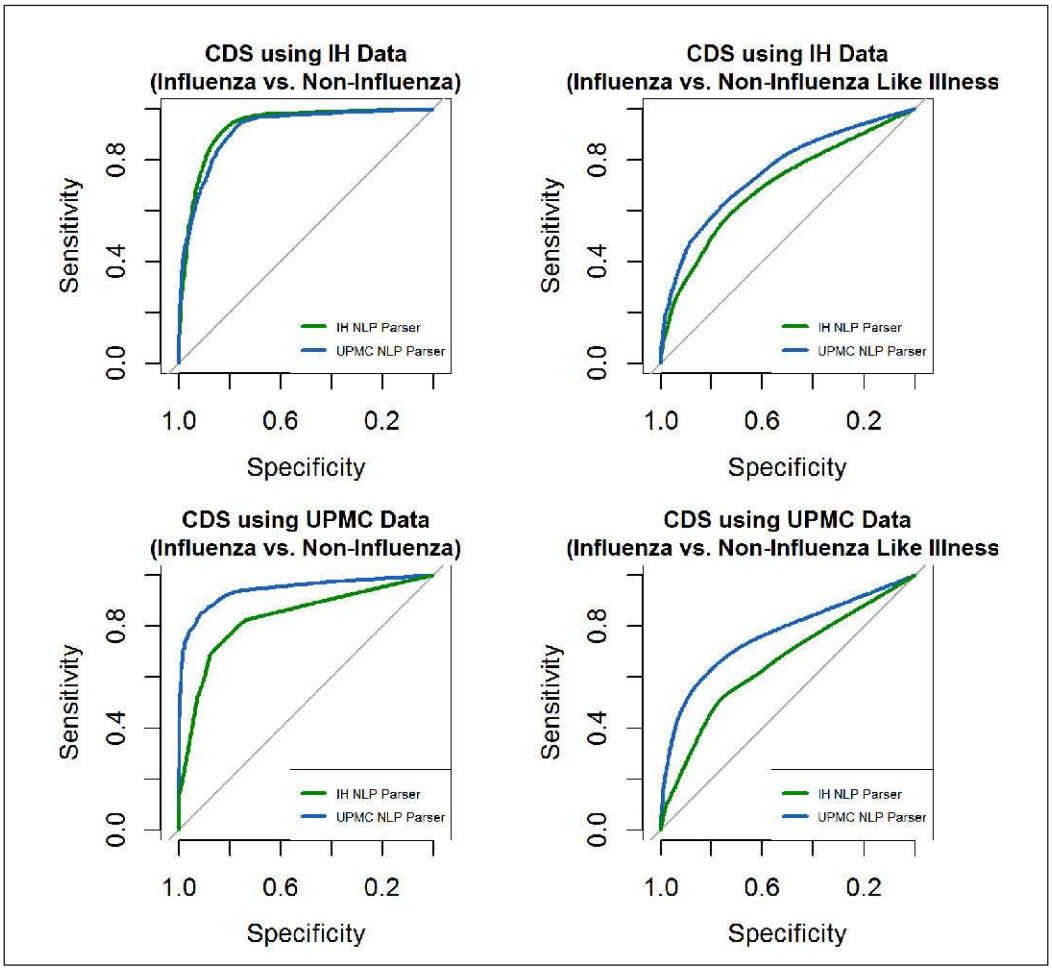

Fig. 3 CDS Performance (AUROC) using different NLP parsers.

Table 1 Top ten annotated clinical findings by institution indicating Influenza

Healthcare System X	Frequency	Healthcare System Y	Frequency
reported fever	580	reported fever	623
other cough	357	wheezing	324
wheezing	349	other cough	310
hypoxemia (sp02 on room air <90%)	278	dyspnea	306
rhonchi	272	tachypnea	239
tachypnea	264	infiltrate	224
vomiting	254	vomiting	221
respiratory distress	250	crackles	206
other abnormal breath sounds	227	abdominal tenderness	201
crackles	220	other abnormal breath sounds	196

Table 2 Summary of NLP Parser Performance

	IH Parser evaluated on IH Corpus	UPMC Parser evaluated on IH Corpus	p Value (IH Parser vs UPMC Parser) on IH Data	UPMC Parser evaluated on UPMC Corpus	IH Parser evaluated on UPMC Corpus	p Value (UPMC Parser vs IH Parser) on UPMC Data
Recall	0.71 (0.70 to 0.72)	0.67 (0.65 to 0.68)	< 0.0001	0.73 (0.71 to 0.74)	0.53 (0.51 to 0.54)	< 0.0001
Recall – Present	0.65 (0.63 to 0.67)	0.63 (0.61 to 0.65)	0.3080	0.68 (0.66 to 0.70)	0.47 (0.44 to 0.49)	< 0.0001
Recall – Absent	0.76 (0.75 to 0.78)	0.69 (0.67 to 0.71)	< 0.0001	0.77 (0.75 to 0.78)	0.58 (0.56 to 0.60)	< 0.0001
Precision	0.75 (0.73 to 0.76)	0.79 (0.78 to 0.80)	< 0.0001	0.80 (0.79 to 0.82)	0.80 (0.78 to 0.81)	0.5188
Precision – Present	0.74 (0.72 to 0.76)	0.70 (0.68 to 0.72)	0.0059	0.79 (0.77 to 0.81)	0.80 (0.78 to 0.82)	0.5612
Precision –Absent	0.75 (0.73 to 0.77)	0.87 (0.86 to 0.89)	< 0.0001	0.81 (0.80 to 0.83)	0.80 (0.78 to 0.81)	0.1639
F1 Score	0.73	0.73		0.76	0.64	
F1 Score – Present	0.69	0.66		0.73	0.59	
F1 Score – Absent	0.75	0.77		0.79	0.67	

Clinical Findings identified in clinical notes are asserted as present or absence.
95% Confidence Intervals in parenthesis.
P-value calculated using X^2 test of two proportions.

Table 3 CDS Performance with Different NLP Parsers with AUC Comparison Tests

	CDS Performance using IH NLP Parseron IH Corpus (Local Performance)	CDS Performance using UPMC NLP Parseron IH Corpus (Portability Performance)	DeLong's test (p-value) H0: Difference between AUROCs = 0 HA: Difference between AUROCs ≠ 0
influenza vs non-influenza	0.932 (0.924 – 0.940)	0.936 (0.928 – 0.944)	0.5176
influenza vs NI-ILI*	0.698 (0.675 – 0.720)	0.748 (0.727 – 0.769)	0.0014
	CDS Performance using UPMC NLP Parser on UPMC Corpus (Local Performance)	**CDS Performance using IH NLP Parser on UPMC Corpus (Portability Performance)**	**DeLong's test (p-value) H0: Difference between AUROCs = 0 HA: Difference between AUROCs ≠ 0**
influenza vs non-influenza	0.954 (0.942 – 0.966)	0.843 (0.820 – 0.866)	< 0.0001
influenza vs NI-ILI*	0.766 (0.735 – 0.796)	0.654 (0.620 – 0.687)	< 0.0001

*NI-ILI: non-influenza influenza-like Illness

References

1. Shaikh AT, Ferland L, Hood-Cree R, Shaffer L, McNabb SJ. Disruptive Innovation Can Prevent the Next Pandemic. Frontiers in public health 2015; 3.
2. Buckeridge DL. Outbreak detection through automated surveillance: a review of the determinants of detection. J Biomed Inform 2007; 40(4): 370-379.
3. Fineberg HV. Pandemic preparedness and response—lessons from the H1N1 influenza of 2009. N Engl J Med 2014; 370(14): 1335-1342.
4. Veenema T, Tõke J. Early detection and surveillance for biopreparedness and emerging infectious diseases. Online journal of issues in nursing 2006; 11(1).
5. Morse SS. Public health surveillance and infectious disease detection. Biosecurity and bioterrorism: biodefense strategy, practice, and science 2012; 10(1): 6–16.
6. Moon S, Leigh J, Woskie L, Checchi F, Dzau V, Fallah M, Fitzgerald G, Garrett L, Gostin L, Heymann DL. Post-Ebola reforms: ample analysis, inadequate action. Bmj 2017; 356: j280.
7. Clemmons NS, Gastanaduy PA, Fiebelkorn AP, Redd SB, Wallace GS, Control CfD, Prevention. Measles—United States, January 4–April 2, 2015. MMWR Morb Mortal Wkly Rep 2015; 64(14): 373-376.
8. Gerbier-Colomban S, Potinet-Pagliaroli V, Metzger M-H. Can epidemic detection systems at the hospital level complement regional surveillance networks: Case study with the influenza epidemic? BMC infectious diseases 2014; 14(1): 381.
9. Control CfD, Prevention. State electronic disease surveillance systems---United States, 2007 and 2010. MMWR: Morbidity and mortality weekly report 2011; 60(41): 1421-1423.
10. Dixon BE, Siegel JA, Oemig TV, Grannis SJ. Towards Interoperability for public health surveillance: experiences from two states. Online journal of public health informatics 2013; 5(1).
11. Gesteland PH, Wagner MM, Chapman WW, Espino JU, Tsui F-C, Gardner RM, Rolfs RT, Dato V, James BC, Haug PJ. Rapid deployment of an electronic disease surveillance system in the state of Utah for the 2002 Olympic winter games. Proc AMIA Symp 2002: 285-289.
12. Centers for Disease Control and Prevention, National Syndromic Surveillance Program (NSSP) – BioSense Platform 2003 [updated March 31, 2016 accessed Apr 2016]. Available from: http://www.cdc.gov/nssp/biosense/index.html.
13. Lombardo J, Burkom H, Elbert E, Magruder S, Lewis SH, Loschen W, Sari J, Sniegoski C, Wojcik R, Pavlin J. A systems overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II). J Urban Health 2013; 80(1): i32-i42.
14. Ferraro J, Haug P, Mynam K, Post H, Li Y, Jephson A, Stoddard G, Vines C, Allen T, Dean N. Performance of a real-time electronic screening tool for pneumonia. Am J Respir Crit Care Med 2012; 185: A5136.
15. Dean NC, Jones BE, Ferraro JP, Vines CG, Haug PJ. Performance and utilization of an emergency department electronic screening tool for pneumonia. JAMA Intern Med 2013; 173(8): 699–701.
16. Moore CR, Farrag A, Ashkin E. Using Natural Language Processing to Extract Abnormal Results From Cancer Screening Reports. J Patient Saf 2014.
17. Ye Y, Tsui F, Wagner M, Espino JU, Li Q. Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. J Am Med Inform Assoc 2014; 21(5): 815-823.
18. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, Gainer VS, Shaw SY, Xia Z, Szolovits P. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. bmj 2015; 350: h1885.
19. Castro VM, Minnier J, Murphy SN, Kohane I, Churchill SE, Gainer V, Cai T, Hoffnagle AG, Dai Y, Block S. Validation of electronic health record phenotyping of bipolar disorder cases and controls. American Journal of Psychiatry 2015; 172(4): 363-372.
20. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. The Oxford University Press; 2013.
21. Chapman WW, Dowling JN, Ivanov O, Gesteland PH, Olszewski R, Espino JU, Wagner MM, editors. Evaluating natural language processing applications applied to outbreak and disease surveillance. Proceedings of 36th symposium on the interface: computing science and statistics; 2004: Citeseer.
22. Chapman WW, Gundlapalli AV, South BR, Dowling JN. Natural language processing for biosurveillance. In: Castillo-Chavez C, Chen H, Lober WB, Thurmond M, Zeng D, editors. Infectious Disease Informatics and Biosurveillance: Springer; 2011. p. 279–310.
23. Dublin S, Baldwin E, Walker RL, Christensen LM, Haug PJ, Jackson ML, Nelson JC, Ferraro J, Carrell D, Chapman WW. Natural Language Processing to identify pneumonia from radiology reports. Pharmacoepidemiol Drug Saf 2013; 22(8): 834-841.

24. Gundlapalli AV, Carter ME, Palmer M, Ginter T, Redd A, Pickard S, Shen S, South B, Divita G, Duvall S. Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among US veterans. AMIA Annu Symp Proc 2013; Nov 16 2013: 537-546.
25. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009; 42(5): 760-772.
26. Elkin PL, Froehling DA, Wahner-Roedler DL, Brown SH, Bailey KR. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. Annals of Internal Medicine 2012; 156(1_Part_1): 11-18.
27. Lippincott T, Séaghdha DÓ, Korhonen A. Exploring subdomain variation in biomedical language. BMC Bioinformatics 2011; 12(1): 1.
28. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc 2011; 18(5): 540-543.
29. Daumé III H. Frustratingly easy domain adaptation. Proc 45th Ann Meeting of the Assoc Compuational Linguistics 2007; 45(1): 256-263.
30. Dredze M, Blitzer J, Talukdar PP, Ganchev K, Graca J, Pereira FC. Frustratingly Hard Domain Adaptation for Dependency Parsing. Conference on Empirical Methods in Natural Language Processing 2007: 1051-1055.
31. Ferraro JP, Daumé H, DuVall SL, Chapman WW, Harkema H, Haug PJ. Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. J Am Med Inform Assoc 2013; 20(5): 931-939.
32. Teixeira PL, Wei W-Q, Cronin RM, Mo H, VanHouten JP, Carroll RJ, LaRose E, Bastarache LA, Rosenbloom ST, Edwards TL. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. Journal of the American Medical Informatics Association 2016: ocw071.
33. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, Pacheco JA, Boomershine CS, Lasko TA, Xu H. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. Journal of the American Medical Informatics Association 2012; 19(e1): e162-e169.
34. Tsui F, Wagner M, Cooper G, Que J, Harkema H, Dowling J, Sriburadej T, Li Q, Espino J, Voorhees R. Probabilistic case detection for disease surveillance using data in electronic medical records. Online J Public Health Inform 2011; 3(3).
35. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. Prentice Hall; 2009. p. 272–319.
36. Bodenreider O. The Unified Medical Language Yystem (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004; 32(suppl 1): D267-D270.
37. Samore MH. Natutal language processing: Can it help detect cases and characterize outbreaks? Advances in Disease Surveillance 2008; 5(59).
38. Pineda AL, Tsui F-C, Visweswaran S, Cooper GF. Detection of patients with influenza syndrome using machine-learning models learned from emergency department reports. Online J Public Health Inform 2013; 5(1).
39. Mehrabi S, Wang Y, Ihrke D, Liu H. Exploring Gaps of Family History Documentation in EHR for Precision Medicine-A Case Study of Familial Hypercholesterolemia Ascertainment. AMIA Summits on Translational Science Proceedings 2016; 2016: 160.
40. Sohn S, Wi C-i, Krusemark EA, Liu H, Ryu E, Wu S, Juhn YJ. Assessment of Asthma Progression Determined by Natural Language Processing to Improve Asthma Care and Research in the Era of Electronic Medical Records. The Journal of Allergy and Clinical Immunology 2017; 139(2): AB100.
41. Liu H, Bielinski SJ, Sohn S, Murphy S, Kavishwar BW, Jonnalagadda SR, Ravikumar KE, Wu ST, Kullo IJ, Chute CG. An information extraction framework for cohort identification using electronic health records. AMIA Jt Summits Transl Sci Proc 2013: 149-153.
42. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association 2010; 17(5): 507-513.
43. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering 2004; 10(3–4): 327-348.
44. Denny JC, Spickard A, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. Journal of the American Medical Informatics Association 2009; 16(6): 806-815.
45. Darwiche A. *Modeling and reasoning with Bayesian networks*: Cambridge University Press; 2009.
46. Ferraro JP, Allen TL, Briggs B, Haug P, Post H, editors. Development and function of a real-time web-based screening system for emergency department patients with occult septic shock. 2008 Annual Meeting – Socity for Academic Emergency Medicine; 2008; Washington, DC.

47. J Leng, S Shen, A Gundlapalli, South B, editors. The Extensible Human Oracle Suite of Tools (eHOST) for Annotation of Clinical Narratives. AMIA Spring Congress; 2010; Phoenix, AZ.
48. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971; 76(5): 378.
49. Cooper G, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine learning. 1992; 9(4): 309-347.
50. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995: 289–300.
51. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes in C: The Art of Scientific Computing. 3rd ed. New York, NY: Cambridge University Press; 2007.
52. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988: 837-845.
53. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association 2010; 17(3): 229-236.
54. Patterson O, Hurdle JF, editors. Document clustering of clinical narratives: a systematic study of clinical sublanguages. AMIA Annu Symp Proc; 2011: Citeseer.
55. Ferraro JP, Daumé H, DuVall SL, Chapman WW, Harkema H, Haug PJ. Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. Journal of the American Medical Informatics Association 2013; 20(5): 931-939.
56. Cooper GF, Villamarin R, Tsui F-CR, Millett N, Espino JU, Wagner MM. A method for detecting and characterizing outbreaks of infectious disease from clinical reports. Journal of biomedical informatics 2015; 53: 15–26.
57. Pineda AL, Ye Y, Visweswaran S, Cooper GF, Wagner MM, Tsui FR. Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. Journal of Biomedical Informatics 2015; 58: 60-69.
58. Shi Y, Sha F. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. Proceedings of International Conference on Machine Learning 2012: 1079-1086.
59. Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW. A theory of learning from different domains. Mach Learn 2010; 79(1–2): 151-175.
60. Blitzer J, Kakade S, Foster DP, editors. Domain adaptation with coupled subspaces. International Conference on Artificial Intelligence and Statistics; 2011.

Appendix A Distribution of Annotated Clinical Findings for Influenza

Influenza Clinical Finding	IH	UPMC	Influenza Clinical Finding	IH	UPMC
Abdominal Pain	92	91	Myalgia	41	91
Abdominal Tenderness	165	201	Nasal Flaring	73	5
Acute Onset of Symptoms	9	24	Nausea	39	121
Anorexia	81	67	Nonproductive Cough	20	28
Apnea	21	9	Other Abnormal Breath Sounds	227	196
Arthralgia	8	5	Other Abnormal X-ray Finding	75	164
Barking Cough	14	5	Other Cough	357	310
Bilateral Acute Conjunctivitis	124	87	Other Pneumonia	113	162
Bronchiolitis	31	16	Paroxysmal Cough	2	4
Bronchitis	4	22	Pharyngitis Diagnosis	18	13
Cervical Lymphadenopathy	110	79	Pharyngitis on Exam	153	130
Chest Pain	34	107	Poor Antipyretics Response	54	23
Chest Wall Retractions	116	51	Poor Feeding	85	96
Chills	10	81	Productive Cough	9	48
Conjunctivitis	1	3	Rales	217	181
Crackles	220	206	Reported Fever	580	623
Croup	19	26	Respiratory Distress	250	192
Cyanosis	100	39	Rhonchi	272	187
Decreased Activity	74	56	Rigor	1	4
Diarrhea	142	145	RSV Lab Testing Only Ordered	18	1
Dyspnea	137	306	RSV Positive Result	45	2
Grunting	70	8	Runny Nose	191	85
Headache	96	126	Seizures	18	19
Hemoptysis	4	12	Sore Throat	77	77
Highest Measured Temperature	189	188	Streptococcus Positive Result	46	10
Hoarseness	9	2	Stridor	47	30
Hypoxemia (Sp02 < 90% on RA)	278	183	Stuffy Nose	127	94
Ill Appearing	85	57	Tachypnea	264	239
Infiltrate	120	224	Toxic Appearance	101	14
Influenza Lab Testing Only Ordered	45	116	Upper Respiratory Infection	51	129
Influenza Positive Result	146	10	Viral Pneumonia	16	1
Influenza-like Illness or URI	210	109	Viral Syndrome	192	136
Lab Ordered (Nasal Swab)	4	62	Vomiting	254	221
Lab Testing 2+ Resp. Pathogens Including Influenza	112	8	Weakness and Fatigue	47	132
Malaise	11	21	Wheezing	349	324