

Evaluating Natural Language Processing Applications Applied to Outbreak and Disease Surveillance

Wendy W. Chapman, PhD¹, John N. Dowling, MD, MS¹, Oleg Ivanov, MD, MPH, MS¹, Per H. Gesteland, MD², Robert T. Olszewski, PhD³ Jeremy U. Espino¹, MD, Michael M. Wagner, MD, PhD¹

¹*RODS Laboratory, Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA.*

²*Intermountain Healthcare, Salt Lake City, UT, USA*

³*Carnegie Speech, Pittsburgh, PA, USA*

Abstract

Much of the pre-existing electronic data that could be harnessed for early outbreak detection is in free-text format. Natural language processing (NLP) techniques may be useful to biosurveillance by classifying and extracting information described in free-text sources. In the Real-time Outbreak and Disease Surveillance laboratory we are developing and evaluating NLP techniques for surveillance of syndromic presentations and specific findings or diseases potentially caused by bioterroristic or naturally-occurring outbreaks. We have implemented a three-stage evaluation process to determine whether NLP techniques are useful for outbreak detection.

First, we are evaluating the technical accuracy of the NLP techniques to answer the question "How well can we classify, extract, or encode relevant information from text?" Second, we are evaluating the diagnostic accuracy of the techniques to answer the question "How well can we diagnose patients of interest using the NLP techniques?" Third, we are evaluating the outcome efficacy of the techniques to answer the question "How well can we detect outbreaks with an NLP-based biosurveillance system?" We give examples from our research for all three levels of evaluation and conclude with suggestions for determining whether NLP is feasible for outbreak and disease surveillance.

Introduction and Background

The appearance of new infectious diseases (e.g., the outbreak of Severe Acute Respiratory Syndrome (SARS) in Asia and Toronto), the reemergence of old infectious diseases (e.g., tuberculosis outbreaks), and the deliberate introduction of infectious diseases through bioterrorism (e.g., October 2001 anthrax attacks) demonstrate the need for surveillance of infectious disease [1]. The United States has not been prepared to deal with biological attacks [2], and biodefense has quickly become a national priority [3]. In response to the need for better biodefense, several research groups have developed electronic surveillance systems [4-13] that monitor a variety of different sources of data including over-the-counter drug sales [14, 15], web-based physician entry of reports [16], 911 calls [17], consumer health hotline telephone calls [18, 19], and ambulatory care visit records [20-24]. Many of the systems monitor pre-existing electronic ED data [25] that typically include date of admission, sex, age, address, coded discharge diagnosis [22, 23, 26], and free-text triage chief complaint [21, 27-30]. Detection algorithms count the number of occurrences of a variable or a combination of variables in a given spatial location over a given time period to look for anomalous patterns [21, 31-33]. If the algorithms detect a significant increase in a given variable, such as the number of patients with a

gastrointestinal illness, the detection algorithms alarm relevant medical and public health officials of a possible outbreak.

The Real-time Outbreak and Disease Surveillance (RODS) system [34] is a biosurveillance system adherent to the CDC's NEDSS standards [35] that was developed in 1999 at the University of Pittsburgh and is currently deployed in four states, including Pennsylvania, Utah, New Jersey, and Ohio. For over 100 hospitals in the four states, RODS collects real-time admission data, including age, sex, zip code, and triage chief complaint. Time-series detection algorithms are applied to the information in the database, and the counts of patients with seven types of syndromic presentations are shown in graphical form on the user interface, shown in Figure 1. The interface also includes a geographic information system that shows counts of syndromic presentations by zip code. If the actual number of patients presenting with gastrointestinal complaints, for instance, exceeds the number expected in a given geographical location over a given time period, RODS' notification subsystem sends an electronic alarm to a team of researchers and public health physicians for possible investigation. RODS software is currently open source [36] and is available for free download at www.health.pitt.edu/rods/sw.

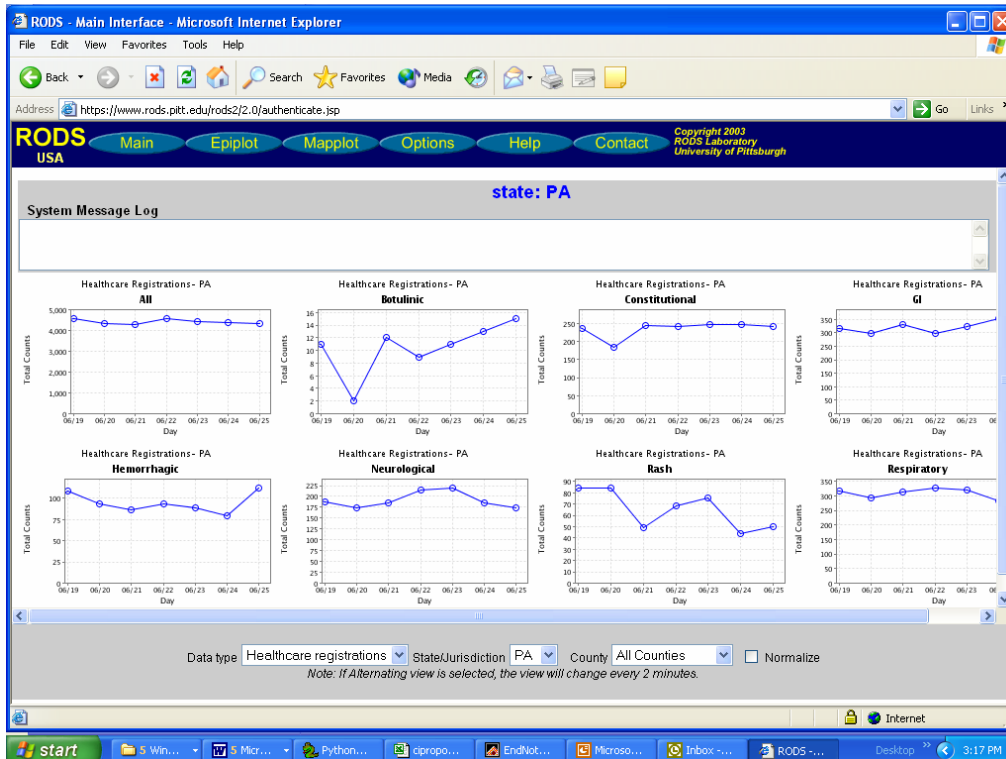


Figure 1. Interface for Real-time Outbreak and Disease (RODS) system, showing syndromic classifications over a one-week period for all admissions in the specified jurisdiction.

Input variables for the detection algorithms in RODS and other biosurveillance systems must be coded data, i.e., data stored in a format that can be interpreted by a computer. For example, a biosurveillance system may monitor the number of patients with Pneumonia, which may indicate a possible outbreak of Influenza, SARS, or inhalational Anthrax. Researchers in medical informatics have developed electronic diagnostic systems integrating multiple sources of data from a patient's medical record to generate a probability that a patient has Pneumonia [37-39]. The variables required to determine the probability of Pneumonia include age and risk factors, vital signs, symptoms and physical findings, laboratory results, blood gas levels, and chest radiograph results. Values for

some of the variables are stored in hospital information systems in coded format; however, variables involving a patient's symptoms and physical findings, such as cough or adventitious respiratory sounds, and the results of a chest radiograph are usually stored as dictated reports in uncoded, free-text format. To use these variables for computerized decision support, the variables must be encoded from the textual reports.

A physician reading the reports could easily determine the correct values for the variables. However, paying physicians to encode reports is impractical. A more feasible solution is applying natural language processing (NLP) techniques to convert the free-text data into an encoded representation that can be used for later inference [40].

Over the last few decades the medical informatics community has actively applied NLP techniques to the medical domain [41, 42]. The Linguistic String Project developed one of the first medical NLP systems that included comprehensive semantic and syntactic knowledge [43-49] that has also been ported to French and German [50-55]. Columbia Presbyterian Medical Center has evaluated and deployed a system called MedLEE [56-60] that extracts clinical information from radiology reports, discharge summaries, visit notes, electrocardiography, echocardiography, and pathology notes. MedLEE has been shown to be as accurate as physicians at extracting clinical concepts from chest radiograph reports [61, 62] and has been evaluated for a variety of applications including detecting patients with suspected tuberculosis [63-65], identifying findings suspicious for breast cancer [66], stroke [67], and community acquired Pneumonia [68], and deriving co morbidities from text [69].

Other medical informatics research groups have also created and evaluated NLP systems for extracting clinical information from medical texts and have shown them to be accurate in limited domains [70-86]. NLP techniques have been used for a variety of applications including quality assessment in radiology [87, 88], identification of structures in radiology images [89, 90], facilitation of structured reporting [72, 91] and order entry [92, 93], and encoding variables required by automated decision support systems such as guidelines [94], diagnostic systems [95], and antibiotic therapy alarms [96].

NLP has only recently been applied to the domain of outbreak and disease surveillance, and most of the research has focused on processing free-text chief complaints recorded in the emergency department [97-102]. We have applied NLP techniques to chief complaints, ED reports, and chest radiograph reports in order to acquire coded variables that may be useful in outbreak detection. To quantify the value of NLP in the domain of biosurveillance, we have adapted a hierarchical model of technology assessment from the domain of medical imaging, described by Thornbury and Fryback [103]. First, we have evaluated the technical accuracy of our NLP techniques to answer the question "How well can we classify, extract, or encode relevant information from text?" Second, we have evaluated the diagnostic accuracy of the techniques to answer the question "How well can we diagnose patients of interest using the NLP techniques?" Third, we have evaluated the outcome efficacy of the techniques to answer the question "How well can we detect outbreaks with an NLP-based biosurveillance system?"

In the Methods section we describe the three levels of evaluation, using the hypothetical example of Pneumonia surveillance as an example. We briefly describe studies we have performed to evaluate NLP technologies for all three levels of evaluation and provide references for details about the studies. In the Results section we provide results from our research for the three levels of evaluation. In the Discussion section, we discuss implications of our findings and suggest three points to consider when appraising the feasibility of applying NLP to the domain of outbreak detection.

Methods

Research in applying NLP technologies to problems of interest in the medical and public health fields is still in the early stages, and NLP systems are only beginning to become accurate enough to be applied to real-world problems. Only a handful of studies have evaluated the impact of NLP technology in healthcare; even fewer have examined its performance in the field of public health. Like other studies involving automated technology, studies in NLP begin by validating the technical accuracy of the technology, and the majority of evaluations of NLP technology in the biomedical domain have focused on this phase of evaluation. Once technical accuracy has been validated, the diagnostic accuracy of the technology can be assessed. In outbreak and disease surveillance, diagnostic accuracy refers to the technology's ability to diagnose or detect specific cases of interest. Finally, summative evaluations addressing outcome efficacy, showing that an NLP-based system can impact the healthcare of a population, can be performed. In the biosurveillance domain, a study validating outcome efficacy would show that an NLP-based detection system can detect epidemics. Figure 2 shows how the three levels of evaluation relate to one another, using a diagnostic system for Pneumonia as an example.

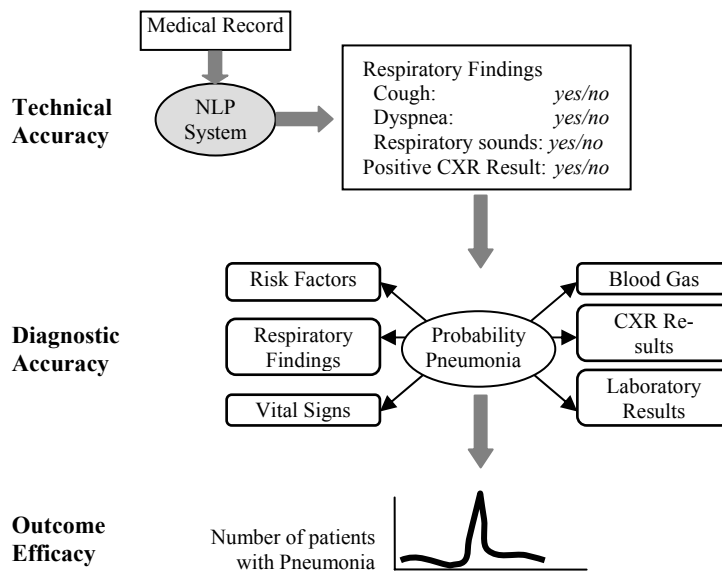


Figure 2. Relationship between the three levels of evaluation for biosurveillance. Evaluations of technical accuracy quantify how well variables and their values are automatically encoded from text. Evaluations of diagnostic accuracy quantify the ability to accurately diagnose a single patient from the variables encoded from text, which may or may not be combined with other variables. Evaluations in outcome efficacy quantify whether the variable being monitored by detection algorithms can detect outbreaks.

The type of evaluation being performed affects both the appropriate reference standard required for calculating performance metrics and the extent to which the results inform us about NLP's contribution to biosurveillance. Evaluations of technical accuracy quantify how well the NLP technology determines the values of relevant variables from text. Therefore, the reference standard must perform the same task and be generated from the same text as the NLP application. Good performance on a technical accuracy evaluation indicates that the NLP application performs the task it was designed to perform. Evaluations of diagnostic accuracy quantify how well assigning values to the variables

contributes to accurately diagnosing individual patients. The reference standard for diagnostic accuracy evaluations is the actual diagnosis of the patient. Good performance on a diagnostic accuracy evaluation indicates that the NLP application can contribute to diagnosing patients. Evaluations of outcome efficacy quantify how well assigning the values to the variables contributes to detection of an outbreak. The reference standard is the presence of an outbreak. Good performance on an outcome efficacy evaluation indicates that the NLP application can contribute to outbreak detection.

Generating a reference standard for evaluation of automated expert systems is difficult, because deciding what the correct response from the system should be and what level of performance is good enough are challenges. Evaluation of biosurveillance systems is sometimes even more challenging for several reasons. First, some of the variables being evaluated have not previously been defined. For example, syndromic definitions, such as gastrointestinal and respiratory, have only recently been explicitly defined for the purpose of disease and outbreak surveillance and have not been externally validated as other case definitions have (e.g., Pneumonia or Influenza). Second, many of the diseases surveillance systems are trying to detect rarely, if ever, occur. For instance, in hospitals in the United States, patients with hemorrhagic or botulinic syndrome and patients with Anthrax or West Nile Virus are rarely seen. Third, the existing reference standard may not be better than the system being tested. For example, the reference standard for existence of an outbreak is currently lab-verified diagnoses and physician reporting of communicable diseases, which may be available later than outbreaks detected by an automated surveillance system and may provide incomplete representation of the extent of the outbreak. Because of the challenges involved in generating a reference standard, evaluations of NLP technology in biosurveillance have begun by answering some of the simpler questions about NLP's contribution to outbreak detection.

Technical Accuracy of NLP in Biosurveillance

The first phase of evaluation for an NLP application should be one of technical accuracy. The question being addressed when measuring the technical accuracy of an NLP application for the domain of outbreak and disease surveillance is *How well does the NLP application determine the values to the variables of interest from text?* For a Pneumonia detector, examples of technical accuracy evaluations include how well the NLP application can determine whether a textual document describes cough, shortness of breath, adventitious respiratory sounds, or radiological evidence of Pneumonia.

We have evaluated the technical accuracy of our ability to classify and encode variables from chief complaints, chest radiograph reports, and emergency department (ED) reports and provide outcome measures for the following NLP tasks:

- (1) **Encoding diarrhea, vomiting [104], and fever [105] from chief complaints.** We calculated the sensitivity, specificity, positive predictive value, and negative predictive value of our ability to use keyword matching to identify chief complaints that indicate diarrhea (e.g., diarrhea, n/v/d, loose stools, etc.), vomiting (e.g., vomiting, vomitting, throwing up, etc.), and fever (e.g., fever, febrile, temp). The reference standard was a physician reading the same chief complaints and determining whether the complaints described any of the variables.
- (2) **Classifying chief complaints into syndromic categories.** Syndromic surveillance is the practice of monitoring any pattern preceding diagnosis for a signal with sufficient probability of an outbreak to warrant further public health response [106]. Grouping cases into syndromes (e.g., respiratory syndrome) rather than into specific diagnoses (e.g., Pneumonia) can provide earlier evidence of infection, because many

diseases in their early phase have overlapping symptoms that may not initially alarm clinicians [107-112]. We developed and evaluated two syndromic classifiers that can classify patients into eight possible syndromic categories based on the patients' chief complaints. For example, a patient with the chief complaint "short of breath" should be classified as respiratory, and a patient with the chief complaint "n/v/d abd pain" should be classified as gastrointestinal. The first classifier is a naïve Bayesian classifier called CoCo [97]. The second classifier is an adaptation of an existing NLP application called MPLUS [113, 114]. We measured the area under the ROC curve (AUC) to determine how accurately the classifiers assigned syndromic categories based on physician gold standard classification of the same chief complaints.

- (3) **Classifying chest radiograph reports consistent with acute bacterial Pneumonia.** In previous studies using medical records from LDS Hospital in Salt Lake City, Utah, we showed that an NLP application called SymText performed similarly to physicians at determining whether chest radiograph reports were consistent with acute bacterial Pneumonia [96]. With this same goal in mind for chest radiograph reports in Pittsburgh, we created a keyword search that accounted for negation and applied it to reports from the University of Pittsburgh Medical Center (UPMC). We compared the keyword search's classification of Pneumonia against that of a physician reading the same chest radiograph reports and calculated sensitivity, specificity, positive predictive value (PVP), and negative predictive value (NPV).
- (4) **Classifying chest radiograph reports describing mediastinal findings consistent with anthrax** [115]. We used the IPS system [116] [117] to classify 79,032 chest radiograph reports based on whether the report described mediastinal lymphadenopathy or widening. We compared the IPS classifications against the baseline of a simple keyword search classifier and calculated sensitivity, specificity, PPV, and NPV for the two classifiers. The reference standard was generated from majority vote of three physicians reading the same reports.
- (5) **Indexing respiratory-related findings from Emergency Department (ED) reports** [118]. We applied an existing indexing application called MetaMap [119] to 28 UPMC ED reports to index individual instances of 71 respiratory-related findings and diseases, such as cough, shortness of breath, pulmonary mass, asthma, Pneumonia, etc. Using a physician as the reference standard, we calculated sensitivity and PPV for MetaMap's ability to identify every instance of the 71 findings within the ED reports.

Diagnostic Accuracy of NLP for Biosurveillance

The question being addressed when measuring the diagnostic accuracy of an NLP application for the domain of outbreak and disease surveillance is *How well does the NLP application diagnose patients from textual data?* For a Pneumonia detector, an evaluation of diagnostic accuracy would determine how well the Pneumonia detector determined whether or not study patients had Pneumonia when compared against a reference standard diagnosis. The reference standard for diagnostic accuracy depends on the finding, syndrome, or disease being diagnosed and may comprise review of textual patient reports or complete medical records, results of laboratory tests, autopsy results, etc.

The NLP application extracts the values of relevant variables from text. Depending on the analysis, the NLP output could be the sole contributor in the system diagnosis, or the variables extracted with NLP may be combined in an expert system like a Pneumonia detector to generate the system diagnoses. Coded variables from other sources, such as microbiology test results or coded admit diagnoses, may also be integrated by the

expert system in generating system diagnoses. Diagnostic accuracy performance is calculated by comparing the diagnoses generated by the reference standard against those generated by the system.

Because evaluations of diagnostic accuracy assess the NLP application's accuracy in relation to the actual clinical state of the patient, diagnostic accuracy evaluations address not only the performance of the NLP technology but also the adequacy of the input text in representing the patient's state. Below we describe examples of diagnostic accuracy evaluations from our research that demonstrate this point.

(1) Classifying patients into syndromic categories based on their chief complaints.

We have performed several studies to determine how well we can classify patients into syndromic categories using only their chief complaints. The difference between technical and diagnostic accuracy evaluations of chief complaint classification is the reference standard: the reference standard for technical accuracy was a physician's classification of the patient from the chief complaint; the reference standard for diagnostic accuracy was diagnosis of the patient either by physician review of the patient's chart or by discharge diagnosis. It is possible to classify a patient correctly according to the chief complaint string but to misclassify the patient based on their actual diagnosis. For example, a patient with a chief complaint of "abdominal pain" may be correctly classified into the gastrointestinal category according to a technical accuracy evaluation. However, the patient's actual diagnosis may be Pneumonia, which is a respiratory syndrome, resulting in an incorrect classification in a diagnostic accuracy evaluation.

We summarized the results of four diagnostic accuracy evaluations of CoCo's ability to classify patients into one of up to seven different syndromes and report the range of sensitivities and specificities for the relevant studies. The first study [99] used physician review of ED reports as the reference standard for identifying patients with acute, infectious gastrointestinal syndrome and compared reference standard classifications against those generated by CoCo's syndromic classification from chief complaints. This study evaluated 585 patients, with 14 positive cases. The second study, based on data described in [120], evaluated CoCo's ability to classify patients into acute, lower respiratory syndrome and compared CoCo's classifications against physician classifications from ED reports [104]. The study evaluated 620 patients, with 30 positive cases. The third study [121] compared CoCo's classifications of patients seen in urgent care facilities in Utah during the Winter Olympic games against two different reference standards: the first reference standard was ICD-9 primary discharge diagnosis; the second was manual classification of patients by Utah Department of Health reviewers, who classified the patients into syndromic categories by review of the chief complaint and the patients' full charts. The Gesteland study examined CoCo's ability to classify 30,094 patients into five syndromic categories, including respiratory, gastrointestinal, neurological, rash, and botulinic, and the number of positive patients ranged from 12 to 2,957, depending on the syndrome. The fourth study [122] evaluated CoCo's classification ability for seven syndromes over a thirteen-year period at UPMC, using primary ICD-9 discharge diagnoses as the reference standard. In this study we evaluated CoCo's classification performance on seven syndromic categories (respiratory, gastrointestinal, neurological, rash, botulinic, constitutional, and hemorrhagic) for 527,228 patients. Positive cases ranged from 1,961 to 34,916, depending on the syndrome.

(2) Diagnosing Fever from Chief Complaints [105]. We calculated the sensitivity and specificity with which we could diagnose patients who were febrile based on keywords in their chief complaints (fever, febrile, temp). The reference standard was

physician determination of fever based on information described in the ED report.

- (3) **Diagnosing Fever from ED Reports** [105]. We developed an NLP application to detect fever from ED reports. The algorithm accounted for negation (e.g., not febrile) with an algorithm called NegEx [123] and for hypothetical findings (e.g., return for fever). We calculated the sensitivity and specificity of the application at diagnosing patients with fever when compared against a reference standard of physician judgment from the same ED report.

Outcome Efficacy of NLP for Biosurveillance

The question being addressed when measuring the outcome efficacy of an NLP application for the domain of outbreak and disease surveillance is *How well does the NLP application contribute to detection of an outbreak?* Two important aspects of evaluating outcome efficacy are predictive performance (i.e., how well the system detects outbreaks) and timeliness (i.e., how soon the system detects an outbreak). For a Pneumonia detector, an outcome efficacy evaluation would measure whether the Pneumonia diagnostic system detected a pneumonic outbreak or whether the system could have detected the outbreak sooner than standard detection techniques.

The first requirement for an outcome efficacy study in outbreak detection is reference standard identification of an outbreak. Outbreaks of respiratory and GI illnesses, such as Influenza, Pneumonia, and Gastroenteritis, occur yearly throughout the country. Outbreaks of other infectious or otherwise concerning diseases, such as Anthrax, West Nile Virus, Hemorrhagic Fever, or SARS, rarely occur in the United States. Once an outbreak is identified, the next requirement for an outcome efficacy evaluation is having access to textual data for an adequate sample of patients living in the geographical area of the outbreak. For instance, if we wanted to evaluate how well our Pneumonia diagnostic system could have detected the 2003 SARS outbreak in Hong Kong had it been deployed at the time, we would need to apply our NLP techniques to relevant textual patient reports generated in Hong Kong in order to extract the values needed for the variables in the Pneumonia detector. Access to personal clinical documents is not easily obtained for research purposes and requires an extraordinary amount of cooperation and trust among researchers and research institutions, hospitals, and local, state, and even federal governments. The two requirements for evaluation of outcome efficacy are not easily attained; therefore, evaluating the contribution of NLP to outcome detection is still in its infant stages.

Studies of technical and diagnostic accuracy of chief complaint syndromic classification were described above, and we have performed one evaluation of the outcome efficacy [100] of chief complaints in detecting pediatric outbreaks. The difference between the previous chief complaint classification evaluations and the outcome efficacy evaluation is the reference standard: the reference standard for the outcome efficacy study was seasonal outbreaks of respiratory and gastrointestinal illnesses. The outcome efficacy evaluation used ICD-9 discharge diagnoses to define retrospective outbreaks of pediatric respiratory and gastrointestinal syndromes using over a five year period (1998-2001) in four contiguous counties in Utah. Sensitivity and specificity of outbreak detection was reported, along with measures of timeliness of detection.

Results

Technical Accuracy of NLP in Biosurveillance

- (1) **Encoding diarrhea, vomiting [104], and fever [105] from chief complaints.** Using keyword matching of chief complaint strings, we were able to identify chief complaints describing diarrhea, vomiting, and fever with 100% accuracy. Every chief complaint considered positive for any of the three variables by the reference standard physician was identified with the keyword searches, and the keyword searches generated no false positives.
- (2) **Classifying chief complaints into syndromic categories.** We developed and evaluated two syndromic classifiers at classifying chief complaints into syndromic categories and show results for the two studies in Figure 3. MPLUS performed with AUC's between 0.95 and 1.0 [114]; CoCo performed with AUC's between 0.78 and 0.97[97].

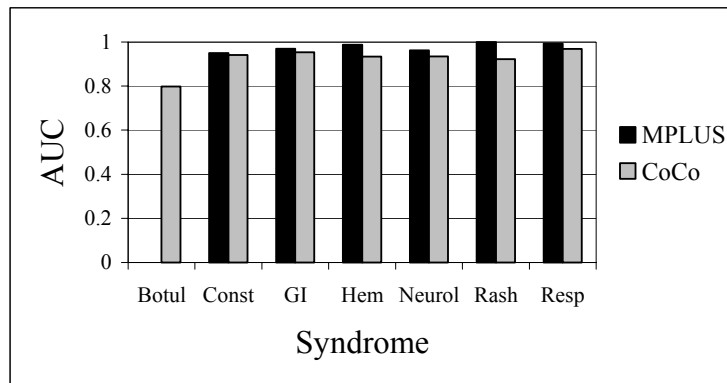


Figure 3. Area under the ROC curve for classifying chief complaint strings into syndromic categories (Botulinic (Botul), Constitutional (Const), Gastrointestinal (GI), Hemorrhagic (Hem), Neurological (Neurol), Rash, and Respiratory (Resp)). MPLUS was tested on a set of 800 chief complaints; CoCo used cross-validation testing on a set of 28,990 chief complaints. There were no botulinic cases in MPLUS' test set.

- (3) **Classifying chest radiograph reports consistent with acute bacterial Pneumonia.** A keyword search with simple negation processing applied to 200 chest radiograph reports identified reports consistent with acute bacterial pneumonia with a sensitivity of 85%, specificity of 96%, PPV of 83%, and NPV of 96%.
- (4) **Classifying chest radiograph reports describing mediastinal findings consistent with anthrax [115].** We compared a simple keyword search against a naïve Bayesian statistical query created using the IPS system on 79,032 chest radiograph reports, of which 1,729 were positive according to the reference standard. Table 1 shows results for both classifiers. We performed a secondary evaluation on the reports by modifying the IPS classifier based on a review of the false negative reports. Sensitivity of the IPS classifier increased to 85.6%, and PPV dropped to 40.9%. Because we used the test set to refine the classifier, results of the secondary evaluation are higher than would be expected on a new test set.

Table 1. Performance of three classifiers at identifying chest radiograph reports describing mediastinal findings consistent with anthrax. Numbers shown are percentages.

	Keyword Search	IPS Model	Refined IPS Model
Sensitivity	43.0	35.1	85.6
Specificity	99.9	99.9	98.8
PPV	96.5	96.5	40.8
NPV	98.6	98.6	99.9

- (5) **Indexing respiratory-related findings from ED reports** [118]. MetaMap indexed respiratory-related findings from 28 ED reports with a sensitivity of 70% and a PPV of 55%. Errors were in large part due to the need to model contextual information in an ED report, such as whether a finding occurred in the past history or at the current visit, and to mistakes from the single physician reference standard.

Diagnostic Accuracy of NLP for Biosurveillance

- (1) **Classifying patients into syndromic categories based on their chief complaints.**

Table 2 summarizes our results from several evaluations in which patients are classified into syndromic categories by CoCo based on their chief complaints. We show a range of the lowest and highest sensitivity and specificity for the evaluations. Overall, about two-thirds of the patients with relevant syndromic presentations were detected by CoCo, with specificities ranging from 90-99%.

Table 2. Diagnostic accuracy evaluations of CoCo’s syndromic classifications. Sensitivities and specificities are shown in percentages.

Syndrome	Number Studies	Reference Standard	Range of Sensitivity	Range of Specificity
Respiratory [104, 121, 122]	5	ICD-9 discharge diagnosis, human chart review	60-77	90-94
Gastrointestinal [99, 121, 122]	4	ICD-9 discharge diagnosis, human chart review	63-74	90-96
Neurological [121, 122]	3	ICD-9 discharge diagnosis, human chart review	68-72	93-95
Rash [121, 122]	3	ICD-9 discharge diagnosis, human chart review	47-60	99
Botulinic [121, 122]	3	ICD-9 discharge diagnosis, human chart review	17-30	99
Hemorrhagic [122]	1	ICD-9 discharge diagnosis	75	98
Constitutional [122]	1	ICD-9 discharge diagnosis	46	97

- (2) **Diagnosing Fever from Chief Complaints** [105]. Using a keyword search on chief complaints, we were able to classify patients according to whether or not the patient actually had a fever with 61% sensitivity and 100% specificity.

- (3) **Diagnosing Fever from ED Reports** [105]. Applying NLP tools to ED reports, we were able to detect patients who were febrile with a sensitivity of 98% and a specificity of 89%.

Outcome Efficacy of NLP for Biosurveillance

We evaluated the ability of syndromic classifications from chief complaints to detect seasonal outbreaks. Figure 4 shows time-series plots from [100] of pediatric chief complaint syndromic classifications against ICD-9 discharge diagnoses for admissions of patients with (a) infectious lower respiratory tract illness due to Pneumonia, Influenza, and Bronchiolitis and (b) infectious gastrointestinal illness due to Rotavirus or other causes of pediatric Gastroenteritis.

Sensitivity and specificity of outbreak detection for respiratory and gastrointestinal outbreaks was 100%. Outbreaks were detected from chief complaints an average of 10.3 days earlier for respiratory and 29 days earlier for gastrointestinal.

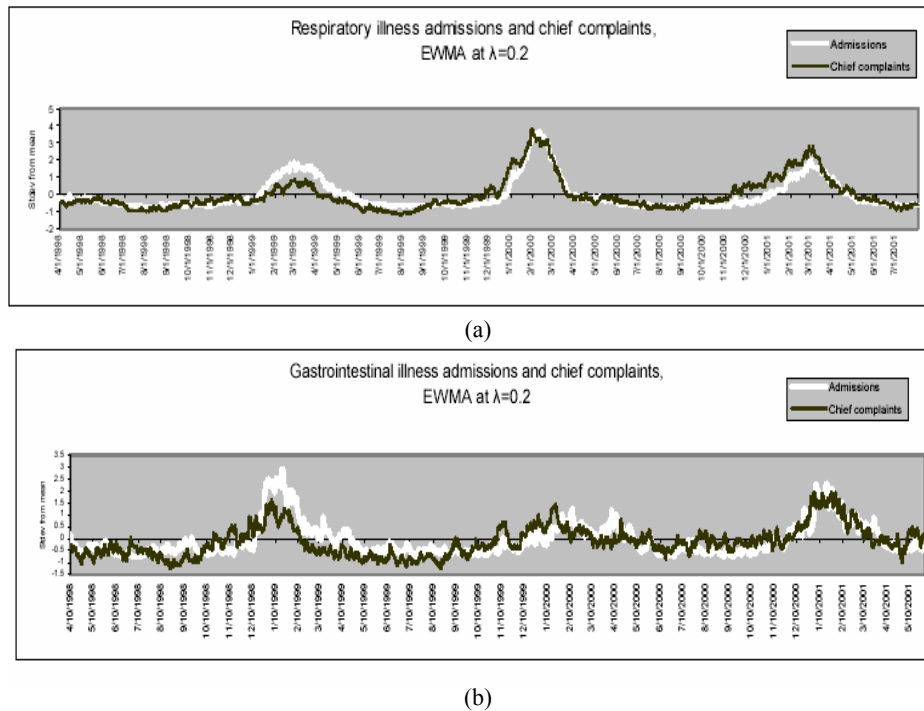


Figure 4. Time series plot of chief complaint syndromic classifications against ICD-9 discharge diagnoses for (a) admissions of patients with Pneumonia, Influenza, and Bronchiolitis and (b) admissions of patients with Rotavirus and other causes of Gastroenteritis.

Discussion

The first step to evaluating a new technology is to measure its technical accuracy. In the domain of biosurveillance, we have applied different types of NLP techniques to chief complaints, chest radiograph reports, and ED reports and have extracted individual findings and classified documents into syndromic or other disease categories.

Our evaluations of technical accuracy suggest that NLP techniques are accurate at identifying single findings from chief complaints. For example, we identified chief complaints describing vomiting, diarrhea, and fever with perfect accuracy. NLP techniques are also very good at classifying chief complaints into syndromic categories and do quite well at classifying chest radiograph reports based on whether the report describes findings consistent with Pneumonia or inhalational Anthrax. Identifying multiple findings from textual reports, such as ED reports, that entail temporality, describe multiple subjects (e.g., patient, physician, other caregivers, and family members), and address findings from multiple anatomic locations is a much more complex task that requires more than the phrase-based or sentence-based techniques we applied in our other studies. To accurately extract multiple findings from ED reports, we need to apply more sophisticated NLP techniques that model not only local information about the finding, but global information about the report as a whole.

The next step in understanding a technology's accuracy in biosurveillance is to evaluate its diagnostic accuracy. Evaluations of diagnostic accuracy compare diagnoses made by the NLP-based system against reference standard diagnoses and convey information about both the accuracy of the NLP system and the quality of the input data. We have shown that CoCo's chief complaint classification can detect about two-thirds of the patients a syndromic surveillance system would ideally detect. This finding has two implications: First, CoCo is accurate enough to detect the majority of relevant patients and second, the majority of the chief complaints of relevant patients reflect the actual syndromic presentation of the patient.

Because evaluations of diagnostic accuracy evaluate not only the NLP application's performance but also the quality of the input data, good technical accuracy does not ensure good diagnostic accuracy. Our studies of fever detection illustrate this point. An evaluation of technical accuracy for identifying chief complaints describing fever showed 100% sensitivity and specificity. However, the fact that we could perfectly identify chief complaints describing fever did not mean we could perfectly identify patients with a fever: when the evaluation was one of diagnostic accuracy so that the reference standard was the patient's actual diagnosis instead of the words in the chief complaint string, our technique maintained 100% specificity, but sensitivity dropped to 61%. Implications of diagnostic accuracy evaluations apply not only to the NLP technique, but also to the data source. Our studies suggest that, in spite of the fact that chief complaints are entered before the patient is examined by a physician and comprise only short phrases, chief complaints are a fairly rich source of information for biosurveillance. The diagnostic accuracy evaluation for fever detection also suggests that to increase sensitivity (e.g., from 61% for chief complaints to 98%), we need to look for information in more detailed clinical records, such as the ED report.

Outbreak efficacy is the most difficult evaluation to perform, requiring collaboration of multiple entities in order to access relevant clinical data and requiring defined outbreaks. Our single study in outbreak detection has reinforced the belief gained from technical and diagnostic accuracy studies that syndromic chief complaint classification can be a powerful source for outbreak detection, at least for respiratory and gastrointestinal outbreaks.

Feasibility of Using NLP for Biosurveillance

Natural language processing techniques are far from perfect. However, the question is not whether the techniques perform perfectly but whether the performance is good enough to contribute to disease and outbreak detection.

We suggest three questions to consider when deciding whether application of NLP techniques to textual data is feasible for disease and outbreak detection: (1) How

complex is the text? The simple phrases in chief complaints are much simpler to understand than complex discourses contained in ED reports. Textual data that require temporal modeling and other more sophisticated techniques to identify values for the variables of interest will be more challenging to process and will be more prone to error; (2) What is the goal of the NLP technique? If the goal is to understand all temporal, anatomic, and diagnostic relations described in the text as well as a physician could, you may be in for a lifetime of work. Extraction of a single variable, such as fever, or encoding temporal, anatomic, and diagnostic relations for a finite set of findings, such as all respiratory findings, is more feasible; (3) Can the detection algorithms that will use the variables extracted with NLP handle noise? Detecting small outbreaks requires more accuracy in the input variables. As an extreme example, automated detection of an outbreak would fail if a single case would be considered a threatening outbreak, which is true of diseases such as inhalational anthrax, and the NLP-based expert system did not correctly detect that case. However, in detecting an outbreak in respiratory syndrome, for example, if the NLP-based expert system only detected two-thirds of the true cases, there may still be enough patients to detect a moderate to large-sized outbreak. In addition, the consistent stream of false positive cases identified by the NLP-based expert system would comprise a noisy baseline that may not prevent the algorithm from detecting a significant increase in respiratory cases but would require a larger increase to detect the outbreak. Consideration of these three questions can help determine the feasibility of using NLP for outbreak and disease surveillance.

Conclusion

NLP techniques can be applied to determine the values of predefined variables that may be useful in detecting outbreaks. The complexity of the textual data being processed and the nature of the variables being used for surveillance determine the feasibility of applying NLP techniques to the problem. Because many of the variables helpful in biosurveillance do not require complete understanding of the text, NLP techniques may successfully extract variables useful for outbreak detection. In fact, our research measuring the technical accuracy, diagnostic accuracy, and outcome efficacy of NLP techniques demonstrates the utility of NLP techniques for a few applications in this new field. More research in NLP techniques and more evaluation studies of the effectiveness of NLP will not only increase our understanding of how to extract information from text but will also help us continue to learn what types of data provide the most timely and accurate information for detecting outbreaks.

Acknowledgments

This work was funded by NLM training grant T15 LM07059, CDC U90/CCU318753-02, DARPA F30602-01-2-0550, AHRQ 1 UO1 HS014683-01, and PA Department of Health ME-01-737.

References

1. M. S. Green, Z. Kaufman. Surveillance for early detection and monitoring of infectious disease outbreaks associated with bioterrorism. *Isr Med Assoc J.* (2002) 4:503-6.
2. D. W. Siegrist. The threat of biological attack: why concern now? *Emerg Infect Dis.* (1999) 5:505-8.
3. D. E. Sanger. Bush plans early warning system for terror. *New York Times* 2002 Feb 6.
4. <http://www.health.pitt.edu/rods/>. Accessed April 16, 2003.
5. <http://www.geis.ha.osd.mil/GEIS/SurveillanceActivities/ESSENCE/ESSENCE.asp>. Accessed April 16, 2003.
6. W. B. Lober, B. T. Karras, M. M. Wagner, J. M. Overhage, A. J. Davidson, H. Fraser, et al. Roundtable on bioterrorism detection: information system-based surveillance. *J Am Med Inform Assoc.* (2002) 9:105-15.
7. <https://secure.cirg.washington.edu/bt2001amia/index.htm>. Accessed April 16, 2003.
8. A. Zelicoff, J. Brillman, D. W. Forslund, J. E. George, S. Zink, S. Koenig, et al. The rapid syndrome validation project (RSVP), a technical paper. *Sandia National Laboratories.* (2001).
9. http://www.nyam.org/events/syndromicconference/2002/posterpdf/buckeridge_poster.pdf. Accessed April 16, 2003.
10. http://www.nyam.org/events/syndromicconference/2002/posterpdf/brillman_poster.jpg. Accessed April 16, 2003.
11. http://www.nyam.org/events/syndromicconference/2002/posterpdf/foldy_poster.pdf. Accessed April 16, 2003.
12. http://www.chip.org/research/biosurv_projects.htm. Accessed April 16, 2003.
13. <http://www.nytimes.com/2003/04/04/nyregion/04WARN.html?ex=1050552000&en=2c63a52eb80102bc&ei=5070>. accessed April 16, 2003.
14. A. Goldenberg, G. Shmueli, R. A. Caruana, S. E. Fienberg. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. *Proceedings of the National Academy of Science.* (2002) 99:5237-5240.
15. http://www.nyam.org/events/syndromicconference/2002/posterpdf/edger_poster.pdf. Accessed April 16, 2003.
16. M. Shannon, J. Burstein, K. Mandl, G. Fleisher. Usage of a web-based decision support tool for bioterrorism detection. *Am J Emerg Med.* (2002) 20:384-5.
17. https://www.stoutsolutions.com/firstwatch/fact_sheet. Accessed April 16, 2003.
18. S. E. Harcourt, G. E. Smith, V. Hollyoak, C. A. Joseph, R. Chaloner, Y. Rehman, et al. Can calls to NHS Direct be used for syndromic surveillance? *Commun Dis Public Health.* (2001) 4:178-82.
19. J. S. Rodman, F. Frost, W. Jakubowski. Using nurse hot line calls for disease surveillance. *Emerg Infect Dis.* (1998) 4:329-32.

20. K. Brinsfield, J. Gunn, M. Barry, V. McKenna, K. Dyer, C. Sulis. Using volume-based surveillance for an outbreak early warning system. *Acad Emerg Med.* (2001) 8:492.
21. B. Y. Reis, K. D. Mandl. Time series modeling for syndromic surveillance. *BMC Med Inform Decis Mak.* (2003) 3:2.
22. R. Lazarus, K. Kleinman, I. Dashevsky, C. Adams, P. Kludt, A. DeMaria, Jr., et al. Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. *Emerg Infect Dis.* (2002) 8:753-60.
23. R. Lazarus, K. P. Kleinman, I. Dashevsky, A. DeMaria, R. Platt. Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. *BMC Public Health.* (2001) 1:9.
24. T. Matsui, H. Takahashi, T. Ohyama, T. Tanaka, K. Kaku, K. Osaka, et al. [An evaluation of syndromic surveillance for the G8 Summit in Miyazaki and Fukuoka, 2000]. *Kansenshogaku Zasshi.* (2002) 76:161-6.
25. http://www.nyam.org/events/syndromicconference/2002/posterpdf/coc/hrane_poster.pdf. Accessed April 16, 2003.
26. M. D. Lewis, J. A. Pavlin, J. L. Mansfield, S. O'Brien, L. G. Boomsma, Y. Elbert, et al. Disease outbreak detection system using syndromic data in the greater Washington DC area. *Am J Prev Med.* (2002) 23:180-6.
27. <http://www.nyam.org/events/syndromicconference/2002/posterpdf/cha/pman.pdf>. Accessed April 16, 2003.
28. C. B. Irvin, P. P. Nouhan, K. Rice. Syndromic analysis of computerized emergency department patients' chief complaints: An opportunity for bioterrorism and influenza surveillance. *Ann Emerg Med.* (2003) 41:447-52.
29. P. H. Gesteland, M. M. Wagner, W. W. Chapman, J. U. Espino, F. Tsui, R. M. Gardner, et al. Rapid deployment of an electronic disease surveillance system in the state of utah for the 2002 olympic winter games. *Proc AMIA Symp.* (2002):285-9.
30. F. C. Tsui, J. U. Espino, M. M. Wagner, P. Gesteland, O. Ivanov, R. Olszewski, et al. Data, Network, and Application: Technical Description of the Utah RODS Winter Olympic Biosurveillance System. *Proc AMIA Symp.* (2002):815-9.
31. W. Wong, A. W. Moore, G. Cooper, M. Wagner. Rule-based anomaly pattern detection for detecting disease outbreaks. *Proc of the 18th National Conference on Artificial Intelligence (AAAI-02).* (2002).
32. G. F. Cooper, D. H. Dash, J. D. Levander, W. K. Wong, W. R. Hogan, M. M. Wagner. Bayesian biosurveillance of disease outbreaks. *Proceedings of the Conference on Uncertainty in Artificial Intelligence.* (2004):(in press).
33. K. Kleinman, R. Lazarus, R. Platt. A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *Am J Epidemiol.* (2004) 159:217-24.
34. F. C. Tsui, J. U. Espino, V. M. Dato, P. H. Gesteland, J. Hutman, M. M. Wagner. Technical description of RODS: a real-time public health surveillance system. *J Am Med Inform Assoc.* (2003) 10:399-408.
35. National Electronic Disease Surveillance System (NEDSS): a standards-based approach to connect public health and clinical medicine. *J Public Health Manag Pract.* (2001) 7:43-50.
36. J. U. Espino, M. M. Wagner, F. C. Tsui, H. D. Su, R. Olszewski, Z. Liu, et al. The RODS open source project for development of syndromic surveillance software. *Proc MEDINFO Symp.* (2004):(in press).

37. D. Aronsky, P. J. Haug. An integrated decision support system for diagnosing and managing patients with community-acquired pneumonia. *Proc AMIA Symp.* (1999):197-201.
38. D. Aronsky, P. J. Haug. Automatic identification of patients eligible for a pneumonia guideline. *Proc AMIA Symp.* (2000):12-6.
39. C. Lagor, D. Aronsky, M. Fiszman, P. J. Haug. Automatic identification of patients eligible for a pneumonia guideline: comparing the diagnostic accuracy of two decision support models. *Medinfo.* (2001) 10:493-7.
40. J. Allen. Natural language understanding. 2nd ed. ed (Redwood City, CA: The Benjamin/Cummings Publishing Company, Inc.; 1995).
41. C. Friedman, G. Hripcsak. Natural language processing and its future in medicine. *Acad Med.* (1999) 74:890-5.
42. P. Spyns. Natural language processing in medicine: an overview. *Methods Inf Med.* (1996) 35:285-301.
43. E. C. Chi, N. Sager, L. J. Tick, M. S. Lyman. Relational data base modelling of free-text medical narrative. *Med Inform (Lond).* (1983) 8:209-23.
44. N. Sager, R. Wong. Developing a database from free-text clinical data. *J Clin Comput.* (1983) 11:184-94.
45. N. Sager, C. Friedman, M. Lyman. Medical language processing: computer management of narrative data (Reading, Massachusetts: Addison Wesley; 1987).
46. R. Grishman, N. Sager, C. Raze, B. Bookchin. The Linguistic String Parser. In: *National Computer Conference* (1973) 427-434.
47. N. Sager. The string parser for scientific literature. In: Rustin R, editor. Natural language processing. New York: Algorithmics Press Inc.; 1973. p. 61-87.
48. N. Sager. Natural language information processing: a computer grammar of English and its applications (Reading, MA: Addison-Wesley; 1981).
49. N. Sager, C. Friedman, E. Chi, C. A. Macleod, S. Chen, Johnson. The analysis and processing of clinical narrative. In: *MEDINFO 86* (1986) 1101-1105.
50. N. Oliver. A sublanguage based medical language processing system for German. New York: New York University; 1992.
51. R. Grishman. Implementation of the string parser of English. In: Rustin R, editor. Natural Language Processing. New York: Algorithmics Press Inc.; 1973. p. 89-109.
52. M. Lyman, N. Sager, E. Chi, L. Tick, N. T. Nhan, Y. Su, et al. Medical language processing for knowledge representation and retrieval. In: *SCAMC 89* (1989) 554-558.
53. N. Nhan, N. Sager, M. Lyman, L. Tick, F. Borst, Y. Su. Medical language processing for knowledge representation and retrieval. In: *SCAMC 89* (1989) 548-553.
54. N. Sager, M. Lyman, L. Tick, F. Borst, N. Nhan, C. Revillard, et al. Adapting a medical language processor from English to French. In: *MEDINFO 89* (1989) 795-799.
55. N. Sager, N. T. Nhan, M. Lyman, L. Tick. Computer analysis of clinical narrative: why how, what, when. In: *BIRA 95* (1995) 22-53.
56. C. Friedman. A broad-coverage natural language processing system. *Proc AMIA Symp.* (2000):270-4.
57. C. Friedman, G. Hripcsak, L. Shagina, H. Liu. Representing information in patient reports using natural language processing and the extensible markup language. *J Am Med Inform Assoc.* (1999) 6:76-87.
58. C. Friedman, P. O. Alderson, J. H. Austin, J. J. Cimino, S. B. Johnson. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc.* (1994) 1:161-74.

59. C. Friedman, G. Hripcsak, I. Shablinsky. An evaluation of natural language processing methodologies. *Proc AMIA Symp.* (1998):855-9.
60. C. Friedman, L. Shagina, Y. Lussier, G. Hripcsak. Automated Encoding of Clinical Documents Based on Natural Language Processing. *J Am Med Inform Assoc.* (2004).
61. G. Hripcsak, C. Friedman, P. O. Alderson, W. DuMouchel, S. B. Johnson, P. D. Clayton. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med.* (1995) 122:681-8.
62. G. Hripcsak, J. H. Austin, P. O. Alderson, C. Friedman. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. *Radiology.* (2002) 224:157-63.
63. N. L. Jain, C. A. Knirsch, C. Friedman, G. Hripcsak. Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. *Proc AMIA Annu Fall Symp.* (1996):542-6.
64. C. A. Knirsch, N. L. Jain, A. Pablos-Mendez, C. Friedman, G. Hripcsak. Respiratory isolation of tuberculosis patients using clinical guidelines and an automated clinical decision support system. *Infect Control Hosp Epidemiol.* (1998) 19:94-100.
65. G. Hripcsak, C. A. Knirsch, N. L. Jain, R. C. Stazesky, Jr., A. Pablos-Mendez, T. Fulmer. A health information network for managing innercity tuberculosis: bridging clinical care, public health, and home care. *Comput Biomed Res.* (1999) 32:67-76.
66. N. L. Jain, C. Friedman. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. *Proc AMIA Annu Fall Symp.* (1997):829-33.
67. J. S. Elkins, C. Friedman, B. Boden-Albala, R. L. Sacco, G. Hripcsak. Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. *Comput Biomed Res.* (2000) 33:1-10.
68. C. Friedman, C. Knirsch, L. Shagina, G. Hripcsak. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. *Proc AMIA Symp.* (1999):256-60.
69. J. H. Chuang, C. Friedman, G. Hripcsak. A comparison of the Charlson comorbidities derived from medical language processing and administrative data. *Proc AMIA Symp.* (2002):160-4.
70. R. K. Taira, S. G. Soderland, R. M. Jakobovits. Automatic structuring of radiology free-text reports. *Radiographics.* (2001) 21:237-45.
71. R. K. Taira, S. G. Soderland. A statistical natural language processor for medical reports. *Proc AMIA Symp.* (1999):970-4.
72. C. A. Morioka, U. Sinha, R. Taira, S. el-Saden, G. Duckwiler, H. Kangaroo. Structured reporting in neuroradiology. *Ann N Y Acad Sci.* (2002) 980:259-66.
73. P. J. Haug, D. L. Ranum, P. R. Frederick. Computerized extraction of coded findings from free-text radiologic reports. Work in progress. *Radiology.* (1990) 174:543-8.
74. P. J. Haug, S. Koehler, L. M. Lau, P. Wang, R. Rocha, S. M. Huff. Experience with a mixed semantic/syntactic parser. *Proc Annu Symp Comput Appl Med Care.* (1995):284-8.
75. P. J. Haug, L. Christensen, M. Gundersen, B. Clemons, S. Koehler, K. Bauer. A natural language parsing system for encoding admitting diagnoses. *Proc AMIA Annu Fall Symp.* (1997):814-8.
76. P. Haug, S. Koehler, L. M. Lau, P. Wang, R. Rocha, S. Huff. A natural language understanding system combining syntactic and semantic techniques. *Proc Annu Symp Comput Appl Med Care.* (1994):247-51.

77. U. Hahn, M. Romacker, S. Schulz. MEDSYNDIKATE-a natural language system for the extraction of medical information from findings reports. *Int J Med Inf.* (2002) 67:63-74.
78. U. Hahn, M. Romacker, S. Schulz. Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. *Pac Symp Biocomput.* (2002):338-49.
79. U. Hahn, M. Romacker, S. Schulz. MEDSYNDIKATE--design considerations for an ontology-based medical text understanding system. *Proc AMIA Symp.* (2000):330-4.
80. R. H. Baud, C. Lovis, P. Ruch, A. M. Rassinoux. A light knowledge model for linguistic applications. *Proc AMIA Symp.* (2001):37-41.
81. R. H. Baud, C. Lovis, P. Ruch, A. M. Rassinoux. A toolset for medical text processing. *Stud Health Technol Inform.* (2000) 77:456-61.
82. M. Romacker, U. Hahn, S. Schulz, R. Klar. Semantic analysis of medical free texts. *Stud Health Technol Inform.* (2000) 77:438-42.
83. W. Ceusters, P. Spyns, G. De Moor. From natural language to formal language: when MultiTALE meets GALEN. *Stud Health Technol Inform.* (1997) 43 Pt A:396-400.
84. W. Ceusters, P. Spyns, G. De Moor. From syntactic-semantic tagging to knowledge discovery in medical texts. *Int J Med Inf.* (1998) 52:149-57.
85. W. Ceusters, J. Rogers, F. Consorti, A. Rossi-Mori. Syntactic-semantic tagging as a mediator between linguistic representations and formal models: an exercise in linking SNOMED to GALEN. *Artif Intell Med.* (1999) 15:5-23.
86. P. Spyns, N. T. Nhan, E. Baert, N. Sager, G. De Moor. Medical language processing applied to extract clinical information from Dutch medical documents. *Medinfo.* (1998) 9 Pt 1:685-9.
87. M. Fiszman, P. J. Haug, P. R. Frederick. Automatic extraction of PIOPED interpretations from ventilation/perfusion lung scan reports. *Proc AMIA Symp.* (1998):860-4.
88. W. W. Chapman, M. Fiszman, P. R. Frederick, B. E. Chapman, P. J. Haug. Quantifying the characteristics of unambiguous chest radiography reports in the context of pneumonia. *Acad Radiol.* (2001) 8:57-66.
89. U. Sinha, R. Taira, H. Kangarloo. Structure localization in brain images: application to relevant image selection. *Proc AMIA Symp.* (2001):622-6.
90. U. Sinha, A. Ton, A. Yaghmai, R. K. Taira, H. Kangarloo. Image content extraction: application to MR images of the brain. *Radiographics.* (2001) 21:535-47.
91. U. Sinha, B. Dai, D. B. Johnson, R. Taira, J. Dionisio, G. Tashima, et al. Interactive software for generation and visualization of structured findings in radiology reports. *AJR Am J Roentgenol.* (2000) 175:609-12.
92. A. B. Wilcox, S. P. Narus, W. A. Bowes, 3rd. Using natural language processing to analyze physician modifications to data entry templates. *Proc AMIA Symp.* (2002):899-903.
93. C. Lovis, M. K. Chapko, D. P. Martin, T. H. Payne, R. H. Baud, P. J. Hoey, et al. Evaluation of a command-line parser-based order entry pathway for the Department of Veterans Affairs electronic patient record. *J Am Med Inform Assoc.* (2001) 8:486-98.
94. M. Fiszman, P. J. Haug. Using medical language processing to support real-time evaluation of pneumonia guidelines. *Proc AMIA Symp.* (2000):235-9.
95. D. Aronsky, M. Fiszman, W. W. Chapman, P. J. Haug. Combining decision support methodologies to diagnose pneumonia. *Proc AMIA Symp.* (2001):12-6.
96. M. Fiszman, W. W. Chapman, D. Aronsky, R. S. Evans, P. J. Haug. Automatic detection of acute bacterial pneumonia from chest X-ray reports. *J Am Med Inform Assoc.* (2000) 7:593-604.

97. R. T. Olszewski. Bayesian classification of triage diagnoses for the early detection of epidemics. In: *Proc FLAIRS Conference* (2003) 412-416.
98. W. W. Chapman, L. Christensen, M. M. Wagner, P. J. Haug, O. Ivanov, J. N. Dowling, et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing. *AI in Med.* (2003) (submitted).
99. O. Ivanov, M. M. Wagner, W. W. Chapman, R. T. Olszewski. Accuracy of three classifiers of acute gastrointestinal syndrome for syndromic surveillance. *Proc AMIA Symp.* (2002):345-9.
100. O. Ivanov, P. Gesteland, W. Hogan, M. B. Mundorff, M. M. Wagner. Detection of Pediatric Respiratory and Gastrointestinal Outbreaks from Free-Text Chief Complaints. *Proc AMIA Annu Fall Symp.* (2003):318-22.
101. D. A. Travers, A. Waller, S. W. Haas, W. B. Lober, C. Beard. Emergency department data for bioterrorism surveillance: electronic data availability, timeliness, sources and standards. *Proc AMIA Symp.* (2003):664-8.
102. D. A. Travers, S. W. Haas. Using nurses' natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department. *J Biomed Inform.* (2003) 36:260-70.
103. J. R. Thornbury, D. G. Fryback. Technology assessment--an American view. *Eur J Radiol.* (1992) 14:147-56.
104. W. W. Chapman, J. N. Dowling, J. U. Espino, M. M. Wagner. Chief Complaint Detection of Syndromic Cases with Broad and Narrow Case Definitions. *Technical Report, CBMI Report Series.* (2004).
105. W. W. Chapman, J. N. Dowling, M. M. Wagner. Fever Detection from Free-text Clinical Records for Biosurveillance. *J Biomed Inform.* (2004) 37:120-7.
106. <http://www.cdc.gov/epo/dphsi/phs/syndromic.htm>. Accessed April 22, 2003.
107. M. J. Kuehnert, T. J. Doyle, H. A. Hill, C. B. Bridges, J. A. Jernigan, P. M. Dull, et al. Clinical features that discriminate inhalational anthrax from other acute respiratory illnesses. *Clin Infect Dis.* (2003) 36:328-36.
108. L. D. Crook, B. Tempest. Plague. A clinical review of 27 cases. *Arch Intern Med.* (1992) 152:1253-6.
109. J. A. Jernigan, D. S. Stephens, D. A. Ashford, C. Omenaca, M. S. Topiel, M. Galbraith, et al. Bioterrorism-related inhalational anthrax: the first 10 cases reported in the United States. *Emerg Infect Dis.* (2001) 7:933-44.
110. Recognition of illness associated with the intentional release of a biologic agent. *MMWR Morb Mortal Wkly Rep.* (2001) 50:893-7.
111. Centers for Disease Control and Prevention.
112. N. Lee, D. Hui, A. Wu, P. Chan, P. Cameron, G. M. Joynt, et al. A Major Outbreak of Severe Acute Respiratory Syndrome in Hong Kong. *N Engl J Med.* (2003).
113. L. Christensen, P. J. Haug, M. Fiszman. MPLUS: a probabilistic medical language understanding system. *Proc Workshop on Natural Language Processing in the Biomedical Domain.* (2002):29-36.
114. W. W. Chapman, L. Christensen, M. M. Wagner, P. J. Haug, O. Ivanov, J. N. Dowling, et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing. *AI in Med.* (2004) (in press).
115. W. W. Chapman, G. F. Cooper, P. Hanbury, B. E. Chapman, L. H. Harrison, M. M. Wagner. Creating A Text Classifier to Detect Radiology Reports Describing Mediastinal Findings Associated with Inhalational Anthrax and Other Disorders. *J Am Med Inform Assoc.* (2003) 10:494-503.
116. G. F. Cooper, B. G. Buchanan, M. Kayaalp, M. Saul, J. K. Vries. Using computer modeling to help identify patient subgroups in clinical data repositories. *Proc AMIA Symp.* (1998):180-4.

117. J. M. Aronis, G. F. Cooper, M. Kayaalp, B. G. Buchanan. Identifying patient subgroups with simple Bayes'. *Proc AMIA Symp.* (1999):658-62.
118. W. W. Chapman, M. Fiszman, J. N. Dowling, B. E. Chapman, T. C. Rindflesch. Identifying respiratory features from emergency department reports for biosurveillance with MetaMap. *Medinfo.* (2004):(in press).
119. A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp.* (2001):17-21.
120. J. U. Espino, M. M. Wagner. Accuracy of ICD-9-coded chief complaints and diagnoses for the detection of acute respiratory illness. *Proc AMIA Symp.* (2001):164-8.
121. P. H. Gesteland, M. M. Wagner, R. M. Gardner, W. W. Chapman, R. T. Rolfs, M. B. Mundorff, et al. Surveillance of syndromes during the Salt Lake 2002 Winter Olympic Games: an evaluation of a naive bayes chief complaint coder. (2004):(in preparation).
122. W. W. Chapman, J. N. Dowling, M. M. Wagner. Syndromic Case Classification from Chief Complaints: a Retrospective Analysis of 527,228 Patients. *Technical Report, CBMI Report Series.* (2004).
123. W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform.* (2001) 34:301-10.