Performance evaluation of a natural language processing tool to extract infectious disease problems

Hannah L Mandel

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington
2013

Committee:

Thomas Payne

Meliha Yetisgen-Yildiz

Robert Harrington

Program Authorized to Offer Degree:

Biomedical and Health Informatics

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

University of Washington

Abstract

Performance evaluation of a natural language processing tool to extract infectious disease problems

Hannah L Mandel

Chair of the Supervisory Committee:
Thomas Payne, Clinical Associate Professor
Biomedical Informatics and Medical Education

Use of a complete problem list can benefit patient care, quality improvement initiatives, and research activities. However, it can be time consuming for physicians to enter the correct encoded problem from a standardized terminology. I evaluated Discern nCode, the natural language processing (NLP) system embedded in Cerner Powerchart at Harborview Medical Center (HMC), for its utility to add Infectious Diseases (ID) problems to the electronic medical record problem list, in comparison with the usual practice of physicians adding problems unaided by NLP. 74 ID consultation notes were annotated by human experts to create gold standard problem lists. NLP-extracted problems and problem list entries were recorded for each note. Recall, precision and f-measure were calculated for nCode and the problem list, and an error analysis was performed to describe false positives and missed concepts. Discern nCode's recall was .65 and precision was .14. Problem list recall was .10 and precision was .43. Many false negatives resulted from partial matches between NLP-extracted and reference standard problems. The majority of false positives were due to inclusion of past medical problems and non-ID problems; nearly 20% of false positives should not have been extracted. Discern nCode had significantly higher recall for ID problems than the problem list. Recommendations are provided for increasing system sensitivity and recall. Overall, nCode could be a useful facilitator of problem entry and result in higher problem list completeness, but recall should be increased.

TABLE OF CONTENTS

INTRODUCTION

In the 1960s, Dr. Lawrence Weed introduced the concept of the problem-oriented medical record as a way to organize medical information and promote diagnostic thinking. He noted:

> "At present the physician has to read the entire record [and] sort the data in his mind if he is to know all the patient's difficulties and the extent to which each has been analyzed. There is no evidence that he does this reliably and consistently; he and others using the record lose their way, and problems get neglected, missed entirely or treated out of context."[1]

Weed envisioned that each medical record would maintain a list all of a patient's problems. He saw this problem list as a dynamic overview of a patient's confirmed diagnoses as well as unexplained findings that have yet to be formalized into a diagnosis. Weed's solutions to this problem are still relevant half a century later, and the idea of a problem list has been widely adopted.

Benefits of the problem list

Today, the problem list is a component of the problem-oriented electronic health record (EHR) and offers a snapshot of the health issues of individual patients. Given the utility of electronic problem lists, their use has been recommended the Institute of Medicine (IOM)[2] and mandated by The Joint Commission on Accreditation of Healthcare Organizations (JCAHO).[3] Meaningful use criteria require hospitals to show that 80% or more of patients seen by eligible providers have a problem coded in the problem list as structured data (or an indication of no problems).[4]

Utilization of the problem list has many purported benefits to patient care. Problem lists provide clinicians with a concise summary of patients' medical histories, lending context for the evaluation of individual complaints[1,5–10] and encouraging a systematic approach to medical problem-solving.[2] Problem lists aid in sign-out and documentation processes.[11–13] and enhance continuity of care[10] as well as follow-up for specific patient problems.[5,8,14] Complete problem lists

also support clinical decision making,[11,13,15,16] quality improvement,[6,17] data mining and clinical research (e.g., through the identification of specific patient populations).[6,13]

Benefits to patient care

Numerous studies have reported on the POMR's benefits to patient care. Simborg et al investigated information factors which affected problem follow-up in ambulatory primary care clinics,[8] and found that problems were more often acknowledged in follow-up visit notes when the records contained a problem list. They concluded that, when compared to problems not on the problem list, recorded problems had a higher follow-up, regardless of the problem, provider specialty, and provider continuity. Hartung at al studied 151 patients with a confirmed diagnosis of heart failure. Of these, 54.4% had heart failure listed in the problem list, and were significantly more likely to be prescribed recommended pharmacotherapy when compared to patients for whom heart failure was not listed as a problem.[5]

Poon et al investigated whether the use of electronic health records was associated with improved quality of care. They used a statewide survey to assess the correlation between specific EHR features (e.g. laboratory and radiology test results, electronic visit notes, reminders, electronic prescribing, medication list, problem list) and Healthcare Effectiveness Data and Information Set (HEDIS) quality measures. Problem list users fared significantly better on HEDIS scores for women's health, depression, colon cancer screening, and cancer prevention measures, scoring 3.3% to 9.6% higher than users without problem list functionality.[9]

Margolis et al found that introduction of problem-oriented medical records into a military clinic increased the number of problems identified and the amount of data collected over the course of new patient visits without increasing visit time,[7] demonstrating benefits of the problem list under the assumption that increased identification of problems leads to higher quality of care.

Additional benefits

Ideally, problem list entries are encoded rather than entered as free text. Because problem list entries map to standardized vocabularies, they can be more useful for research and for EHR functionalities such as clinical decision support, quality assurance, generation of billing codes, and public health reporting.[18–20] For example, Jao et al designed a clinical decision support system that compared ordered medications with patients' problem list entries in order to identify mismatches. This system lead to improvements in completeness of the medication list as well as the problem list,[21] demonstrating that the problem list can be used for quality assurance activities which may result in increased patient safety.

Advantages of problem list data

It is difficult and time consuming to manually extract data from unstructured notes. As mentioned, information locked in narrative text cannot be used by automated systems to improve care, facilitate research, or be integrated with other data sources.[19,22] Therefore, maintaining a complete, structured problem list is superior to leaving diagnoses embedded in clinical notes, and also superior to reliance on proxy methods such as billing codes, which are notoriously unreliable.

Several studies have looked at the use of administrative codes to identify adverse events, and have found that they severely underestimate their prevalence in hospitals. Meddings et al compared frequency of use of the correct ICD-9 code for catheter-associated urinary tract infections (CA-UTIs) to the gold standard of physician-reviewed patient medical records.[23] They found that the gold standard identified 45% of the patient UTI cases as being CA-UTIs, but that none of these patients had been assigned the corresponding ICD-9 code. In a study comparing three methods to detect adverse events at three US hospitals, Classen et al found that Patient Safety Indicators (based on claims data) detected only 9% of all adverse events, though it is

unclear whether this is due to lack of assignment by coders or an insufficiency of claims vocabularies.[24] Despite this, International Classification of Diseases (IDC) 9 codes are currently the most widely used sources of clinical information in the US.[25,26] It is clear that problem list data could greatly enhance quality assurance activities. The accuracy of patient data is increasingly important as, by 2015, hospitals nationwide will be compared based on their rates of hospital-acquired events.[27]

Barriers to problem list use

The problem list is not effective if it is not maintained, and problem lists are often underutilized, in part because they are viewed as incomplete, inaccurate or not up to date.[28–30] Szeto et al showed that, in a general medicine clinic, only 49% of patients with coronary artery disease, 42% with benign prostatic hyperplasia, and 81% with diabetes had these diagnoses documented in the problem list.[28] In addition, several other reviews at U.S. hospitals[20,31,32] have found that problem lists are extremely underutilized and have an abundance of free text entries, with the University of Illinois showing that 47% of entries were free text instead of coded.[32]

Barriers to the use of problem lists may exist because of the lack of guidelines or a common approach to problem entry. Clinicians may be uncertain as to who is accountable for entering problems and responsible for keeping the list current.[30,33] They may also be confused about what qualifies as a "problem," what problems are worthy of being entered into the list, and at what level of precision is appropriate.[34,35]

The task of identifying the correct code for a problem can present an additional barrier. The U.S. Department of Health and Human Services specifies that problem list entries must be coded using ICD-9-CM or the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT).[4] These codes can be entered manually, and entry is often facilitated by use of a pick list or search

engine.[36] However, providers find it challenging to integrate problem entry with their workflow and to identify proper codes, as codes (particularly ICD-9 codes) may not be intuitive.[34]

Meanwhile, the study of NLP to automate problem identification has evolved in an effort to streamline problem entry, with the ultimate goal of improving problem list accuracy and completeness. NLP has been used to facilitate problem entry by providers[31,36–38] and to automatically add terms to the problem list.[11,21]

Background: natural language processing

The application of NLP to extract data from clinical narrative text has been an active area of research since the 1960s. In the past decade, NLP has been used for a variety of purposes including pharmacovigilance;[39] extracting medication information from discharge summaries;[40,41] extracting clinical information (i.e. infection with pneumonia) from radiology reports;[42–44] medical record de-identification;[45] and identification of infectious disease symptoms in emergency department patients.[46]

Clinical natural language processing challenges

Research applying NLP to clinical narrative text must consider unique domain-specific challenges:

Structure and content

The structure and content of clinical text is unique. Clinical notes favor brevity, and may not use full sentences or follow grammar rules. Notes are also prone to spelling errors, which can be numerous in medical documents not meant for publication, especially if typed rather than dictated and professionally transcribed.[47] Additionally, the structure of clinical documents can

make processing difficult, as they often include diagrams, lists of vital signs, and non-textual clinical information.[48]

Terminology

Biomedical terms are prone to a large amount of ambiguity. Polysemy refers to one word having multiple meanings, and linguistic variation refers to many words having the same meaning; both of these contribute to ambiguity in the medical record.[49] Moreover, the interpretation of terms within clinical text is highly contextual.

Abbreviations

Clinical text is rife with forms of shorthand.[48] According to Liu et al, "abbreviations [...] have multiple full forms, for instance, the abbreviation APC refers to *activated protein c, adenomatosis polyposis coli, adenomatous polyposis coli, antigen presenting cell, aerobic plate count, advanced pancreatic cancer, age period cohort, alfalfa protein concentrated, allophycocyanin, anaphase promoting complex, anoxic preconditioning, anterior piriform cortex, antibody producing cells, and atrial premature complex*, etc. in the MEDLINE abstracts."[50] More than one third of abbreviations in the Unified Medical Language System (UMLS) with fewer than seven characters are overloaded.[50]

Negation

Chapman et al found that negations were common in clinical reports, and most frequent in radiology reports. UMLS phrases in clinical reports were negated approximately 40-80% of the time with the most common negation terms being "no," "without," "not," and "denies/denied."[51] Negation detection is important for tasks such as classification, where the system must identify whether certain concepts are absent or present.

Ambiguity of clinical information

Clinical text contains layers of temporality, implication and uncertainty which are simple for people to discern but prove difficult for NLP to understand and derive true meaning.[48,52] For example, physicians can use the structure of a document and context clues to determine whether problems are current. If a problem is listed under "past medical history," or a patient is described as having "history of" a problem, it is evident that the problem occurred in the past.

Clinical notes often contain an implicit diagnosis rather than an explicit one, and a concept's presence does not necessarily mean that it qualifies as a current diagnosis. It may be part of a differential diagnosis, a past but resolved ailment, or family history (i.e. "family history of type 2 diabetes mellitus").

The differential diagnosis can include many expressions of uncertainty. For example, it could include statements such as "we were unable to rule out pneumonia," "we are trying to rule out pneumonia," "the patient may have pneumonia," "possible pneumonia."[53]

Natural language processing methods

NLP methods are generally symbolic or statistical. Symbolic (or grammatical) techniques utilize the structure of language and the meaning of the words to encode free text. Statistical approaches are trained on large datasets to use the frequency distribution of words in order to encode text.[53] Many current efforts at clinical NLP use a hybrid approach combining statistical and rule-based approaches.

Symbolic techniques exploit the characteristics of language (i.e. semantics, syntax and parts of speech). Syntax refers to the arrangement of words within a sentence or phrase. Syntactic methods involve tagging individual words with their part of speech and breaking down sentences into their constituent parts (known as parsing), creating a tree which identifies the

relationships between parts of a sentence. Parsers use a rule-based or statistical grammar to determine the structure of a sentence.

The syntax of a sentence can help to determine the semantic relationship between words. However, the processor must also use semantic relationships of individual words to determine overall meaning. For example, the Semantic Network of the UMLS categorizes Metathesaurus terms with semantic types (i.e., the heart is an anatomical structure) and further defines relationships between semantic types (i.e. the aorta is part of the heart).

Symbolic processing usually begins with a preprocessing step, section boundary detection, sentence boundary detection, and tokenization.[54] The end result of these steps is a phrase of interest which has been separated into words. This is followed by a phrase parsing step, in which parts of the sentence are tagged with their parts of speech and then mapped to concepts. For example, “the margins of resection” could be mapped to an ontology such as the UMLS, and the concept code for “resection margin” would be identified.[54] This process of mapping to a controlled vocabulary has the benefit of reducing the complexity of the text. Once all phrases have been mapped, their concepts are strung together in a post-coordination step to make a determination.

Statistical NLP deviates from the above method and uses a statistical approach such as supervised classification approaches (e.g., support vector machines). Wendy Chapman uses the example of the word “discharge” to illustrate that a statistical learning technique could differentiate between its two meanings; she shows that nearby words such as “prescription” and “home” would indicate that a patient is being discharged from the hospital, while nearby words such as “rashes” and “wound” would indicate that the discharge is a bodily fluid.[53] Machine

learning techniques require a large training set, which may be expensive and time-intensive to generate, but can be more easily adapted to other domains.[48]

Natural language processing and problem extraction

There have been numerous strides towards the use of NLP to classify conditions or diagnoses from clinical text. Hripscak et al reported on the use of MedLEE to identify six conditions (e.g., congestive heart failure) from 200 chest radiograph reports. The processor's performance, as measured by sensitivity and specificity, was equivalent to that of physicians and better than laypeople.[22] System sensitivity was 81% and specificity was 98%. The physicians' average sensitivity was 85%, with a specificity of 98%.

Chute et al performed an evaluation of NLP algorithms to automatically classify diagnoses as Hospital International Classification of Disease Adaptation 2 (HICDA-2) codes at the Mayo Clinic in Rochester, MN. They found that automated coding was faster, but would necessitate human verification.[55] Building on work at the Mayo Clinic, Pakhamov et al created an NLP system which incorporated machine learning to classify diagnoses. Free text from clinical notes was processed, and diagnoses were classified using codes within the HICDA. The authors determined that nearly 50% of EHR problem list entries could be automatically classified with precision and recall above 98%. This system was implemented, replacing manual coders at the Mayo Clinic.[56]

Extensive research on this topic has also been conducted at the LDS Hospital and University of Utah in Salt Lake City, Utah. The Medical Informatics group at the University of Utah has a history of NLP application development, including the development of SymText (Symbolic Text processor),[57] a hybrid NLP application using syntactic and probabilistic semantic methods.[36] Gunderson et al developed the Automatic Admit Diagnosis Encoding System (AADES)[18,19] to

code admission diagnoses from free-text documents using SymText's semantic and syntactic techniques. They found that the system was 76% accurate. Out of all coded cases, accuracy was equivalent to manual coding.[18]

NLP has also been used specifically to aid in problem entry. Meystre and Haug hypothesized that using NLP to generate potential medical problems would improve problem list completeness. They identified a list of 80 medical problems which were common to the LDS Hospital, and used the Automated Problem List (APL) system to extract these problems from free-text clinical documents. The APL consists of a background application which uses NLP to extract and store the identified medical problems, and a problem list management application integrated with the EHR which allows for editing and creation of problems as well as links to relevant internet sources. In earlier publications, Meystre and Haug used MPLUS, another hybrid NLP iteration, within the background application. They later used the Java version of MetaMap (known as MMTx) in combination with negation detection based on NegEx[58] to study the application of NLP to problem entry. In a randomized clinical trial, they found that their system significantly increased the sensitivity of problem lists of ICU patients from 9 to 41%.[31]

William Long recently evaluated the use of NLP to extract medical concepts from discharge summaries with the goal of providing physicians with concepts to add to the problem list. The system was optimized to maximize sensitivity and, though it had a high false positive rate, extracted 93% of desired concepts.[59]

Solti et al extended the work of Haug and Meystre at the University of Washington.[60] The authors followed similar methods, using the UMLS and MetaMap to construct a natural language processer without negation capabilities in order to automatically populate the problem list. The application had 88% sensitivity and 66% precision, and the authors noted that

ambiguity was responsible for a loss of performance. Currently, our institution has implemented Cerner's Discern nCode to automate problem entry. Discern nCode is based on Peter Elkin's work on the Multi-threaded Clinical Vocabulary Server (MCVS).

Multi-threaded Clinical Vocabulary Server

Elkin et al introduced the notion of Concept-Based Indexing[26,61] and hypothesized that a more detailed, granular terminology such as SNOMED would offer superior coding of clinical records due to its support of compositionality, by which atomic concepts can be combined, resulting in greater specificity (e.g. *history* of *severe* COPD). The ability to encode compositional expressions means that problem list entries, or diagnostic codes, could be more accurate and specific.[62] The MCVS was developed to encode medical problems from narrative text, and incorporated the capability to form compositional expressions and negation detection.[63] An early version of the negation detection for MCVS was evaluated using outpatient clinical records from Johns Hopkins Medical School; the sensitivity for negated concepts was 97% and the specificity was 99%.[64]

Further work using the MCVS and concept-based indexing has shown promise for information extraction. Brown et al found that the MCVS could be used to identify quality indicators present in spine disability examinations at the Veterans' Health Administration.[65] To accomplish this, the MCVS first divides the examination into sections based on their clinical content (e.g. patient history). Strings are normalized using the UMLS's normalization tool (Norm). Norm performs a series of lexical variant generation transformations; in this process, words are converted into their uninflected forms (e.g. "diseases" becomes "disease") and stripped of possession, casing, and punctuation. After normalization, sentences are broken into words and phrases, which are mapped to SNOMED CT and assembled into compositional expressions which account for negation and uncertainty. A rule-based evaluation engine then processes the resulting

expressions; Boolean rules are manually created, iteratively modified and assessed on a set of training data. The authors found that this method had a sensitivity of 87% and specificity of 71% for quality indicators in spine disability examination reports, which have a structure similar to other medical notes. An extension of this study was subsequently performed using a commercial version of the MCVS based on the research version and yielded similar results.[66]

Elkin et al also evaluated the ability of the MCVS to classify pneumonia as positive, uncertain, or negative from radiological reports.[42] The system had a sensitivity of 100% and specificity of 90.3%. The errors which resulted in reduced specificity were due to the system's classification of pneumonia assertions as positive despite being listed within differential diagnosis sections.

Motivation for system evaluation

NLP systems are evolving and have the potential to eliminate human errors in problem entry, increase the speed and timeliness of entering coded problems, and make problem lists more complete. Incorporation of NLP could also improve compliance with meaningful use requirements at our institution. I wanted to know how well Discern nCode identified terms for problem entry, and how this compared to problem entry performed by physicians. The goals of this summative evaluation of Discern nCode are:

1. To test the hypothesis that Cerner Discern nCode can perform as well as physicians at adding problems from clinical narrative text to the problem list. This involves an evaluation of Discern nCode's performance at extracting problems from notes, as well as an assessment of the frequency of post-visit problem entry by physicians.
2. To describe and categorize the reasons for mismatches between Discern nCode extracted problems and the content of the corresponding inpatient note.

METHODS

This study was approved by the University of Washington (UW) Human Subjects Division.

Study setting

Harborview Medical Center (HMC) is located in Seattle, WA and is part of UW Medicine. Harborview has 413 licensed beds and treats a large number of uninsured, poor, and homeless patients, as well as the mentally ill and drug abusers. It has a large Infectious Diseases (ID) section which treats a wide variety of rare and complicated infections. The ID clinic sees patients five days per week and provided over 1,000 visits in 2011. As part of a small grants application awarded by the Institute of Translational Health Sciences, the ID section has worked to create a database of provider-assigned problem list entries from a pick-list of infectious diseases and organisms for quality improvement initiatives and research, providing a data source for this study.

Information Technology at HMC

Use of the electronic health record by the UW Medicine began in 2003 with the implementation of a commercial EHR.[67] UW Medicine now has a wealth of clinical information systems. The Online Record of Clinical Activity (ORCA) is the inpatient Electronic Health Record at Harborview, and uses a Cerner PowerChart application for charting. Problem list entries, as part of the permanent medical record, are entered into ORCA. UW Medicine also uses the Computerized Rounding and Sign Out System (UWCores), which has been shown to improve patient care;[68] problem list entries are also entered into UWCores but do not become part of the medical record.

The electronic problem list is accessible from the Diagnoses & Problems tab within ORCA. Problems can be added manually through a search function. Some departments at HMC, such as

Cardiology and ID, have created pick-lists of the most common syndromes and/or organisms to aid in problem entry.

NLP is available to semi-automate problem entry at UW Medicine hospitals and the ORCA application offers access to Discern nCode, which extracts SNOMED CT, ICD-9, CPT and E&M codes from clinical text. After Discern nCode processes a selected note, it provides a list of suggested problems to an mPage display designed by UW Medicine within a PowerChart tab. Any problem displayed on that mPage that is clicked on by the physician will be automatically added to the patient's problem list, circumventing the process of manually searching for and entering the correct problem and code.

UW Medicine has partnered with Microsoft to use Microsoft Amalga Unified Intelligence System as a data integration platform and clinical data repository. Amalga aggregates data across UW Medicine's clinical information systems and is currently used for quality improvement and translational research initiatives. Amalga currently serves as a clinical data repository for UW Medicine.

Infectious Disease workflow

The Infectious Disease Consult Service at HMC performs inpatient consultations, generally rounding throughout the day, with teaching and attending rounds in the afternoon. Consultations result in a note in ORCA, which follow an infectious disease note initial consultation or progress note template. These notes outline chief concerns, history of the present illness, physical exam, vitals or review of systems, medications, laboratory or radiological studies, past medical history, family and social history, and a section on problems, assessment and plan.

In practice, ID physicians enter the final ID diagnosis and organism (e.g. *aortic valve endocarditis* and *Staphylococcus*) into the problem list if both are known. If there is not a final diagnosis at the time of the initial consultation, as in the case of a fever with no obvious infection, *fever* may be entered into the problem list and later replaced with the final diagnosis; however, organisms are only entered into the problem list if they are laboratory-confirmed. Physicians are encouraged to select problems from the ID pick list. Documentation is not necessarily performed the same day as the consultation.

Data collection

Note identification

ORCA and Amalga were used to obtain the data for this study. Amalga was queried to identify HMC patients who were seen in consultation by a member of the inpatient ID consult team after November 2012, had an associated note (entitled *Infectious Diseases Inpt Record* or *Consultation Inpt*), and had at least one problem recorded in their problem list. The resulting dataset included patients' HMC medical record numbers, ID note date of entry, the physician responsible for the note, and the number of problems in the patient's problem list. For patients with multiple ID consult notes, the initial consultation note was chosen for analysis unless ID problems were more conclusively established in a follow-up note. Notes were excluded from this analysis if they conclusively determined that there were no infectious disease problems, as these notes could not be used to create gold standard problem lists. In total, 74 notes belonging to unique patients were identified.

Using ORCA, the note for each patient was located. Notes were manually stripped of all identifiers in accordance with the HIPAA Privacy Rules [45 CFR 164.514].[69] Two physicians annotated the notes to create reference lists of problems against which problem list entries and

NLP identified problems were compared. The patients' problem list entries were documented along with Discern nCode extracted problems for each note. Some text from a typical note is below:

> "He then returned to the ED and was found to have rapidly progressive erythema across his L abdomen/flank and was taken to the OR immediately, where he was found to have a significant amount of dead bowel. His abdominal wall began to become increasingly involved over hours and ultimately had nearly all of his anterior and L abdomnial wall needed to be resected and he was also given a loop ileostomy. Intraop cultures from TGH showed C. albicans and various strep spp."

Creating the gold standard lists

Two UW Medicine physicians were recruited to review each note and identify the terms which they would consider to be potential infectious disease problem list entries; the resulting lists were considered to be the gold standard for each note. Because this was an evaluation of Discern nCode's ability to identify the same problems that ID physicians would record in the problem list after a consultation, only infectious disease syndromes and organisms were annotated.

For guidance about what problems would be relevant to infectious disease, annotators were given a copy of all terms within the ID pick list. For the purpose of this evaluation, a problem was defined as 1) a syndrome that a) has a diagnosis or b) is confirmed despite unknown etiology (e.g., fever), or 2) a laboratory-confirmed organism, and notably important in consideration of the patient's care or treatment. Reviewers were instructed to annotate problems that were ongoing or new at the time the note was written, but not if they were resolved aspects of a patient's medical history. Problems were not annotated if they were possible but not confirmed (i.e., in the differential diagnosis), or if they had been negated (i.e., "there was no evidence of").

These rules were iteratively refined by the annotators and investigator. Five notes were double-coded by both reviewers. P_{pos}, the inter-annotator agreement, was an average of 0.52 per note.[70]

Some of the reasons for low inter-annotator agreement were misunderstanding of the clinical situation and lack of an objective rule regarding how "important" problems needed to be. Rothschild et al found that differences in clinical styles and ambiguity in fuzzy diagnoses can lead to poor inter-annotator agreement in problem list coding.[71] After meeting to resolve conflicts and clarify annotation guidelines, five more notes were double-coded. The inter-annotator agreement was 0.67. After meeting again to resolve conflicts, consensus was reached and the remainder of the notes was coded independently by either AW or DM.

Comparison to the gold standard

The gold standard list for each note was compared to the list of NLP-extracted problems to identify false positives, false negatives, and true positives (Table 1). True negatives were not assessable because the corpus of potential infectious disease problems was not possible to quantify. Because Discern nCode was designed to extract all problems, not just those in the ID domain, I expected a high number of false positives.

Table 1. Discern nCode evaluation classifications

		Gold Standard list	
		Present	Absent
Discern nCode	Present	TP*	FP
	Absent	FN	N/A

*Problems had to be an exact conceptual match within the UMLS to be considered a true positive. This was assessed by HM using the UMLS Terminology Services Metathesaurus Browser (uts.nlm.nih.gov//metathesaurus.html). This definition of a true match has been used in prior evaluation studies.[63]

The gold standard lists were also compared to the patients' current problem lists (Table 2). Post-visit problem entry was the primary measure of interest. Because there was no straightforward

way of associating notes and problem entries, and they may not necessarily be entered on the same day, only problem list entries recorded within a one-week window around the date of the note were considered. Problems automatically entered by the system, instead of by a physician, were not included. Because this approach may penalize physicians who did not enter a problem because it was already present in the problem list, the entire problem list was also evaluated.

Table 2. Problem list evaluation classifications

		Gold Standard list	
		Present	Absent
Medical Record	Present	TP*	FP
	Absent	FN	N/A

*True positives were exact UMLS concept matches, as assessed by the investigator (HM). It was not necessary that a problem list entry be entered by the same physician responsible for the visit note, as ID physicians often collaborate and may share documentation tasks.

Data Analysis

Inter-Annotator agreement

While the Kappa coefficient is a traditional measure of agreement used in NLP evaluations, it would be necessary to know the number of times physicians agree that a term should not be annotated.[72] Because of the difficulty in ascertaining this, I employed methods described by Hripscak and Rothschild in which they suggest that calculating the average f-measure between pairs of raters approaches the Kappa statistic when the number of true negatives is unknowable.[70]

Performance evaluation metrics

Recall (also known as sensitivity), precision and f-measure were used to assess the performance of NLP. NLP with perfect precision would not identify any problems that were not in the gold standard, while NLP with perfect recall would identify all the problems in the gold standard. In

this study, a higher recall has greater value than a high precision, as it would be easy for physicians to discard false positives from a list of potential problem list entries by clicking only on true positives, while physicians must resort to manual problem entry if the list is missing an applicable concept.

For each note, the number of true positives, false negatives, and false positives were scored and totaled. The recall and precision were calculated from the totals (Table 3). From the recall and precision, the f-measure was calculated.

Table 3. Performance measure equations

Recall	$\frac{TP}{TP + FN}$	Proportion of gold standard problems identified; sensitivity.
Precision	$\frac{TP}{TP + FP}$	Proportion of extracted or problem list problems in the gold standard.
F-measure	$\frac{(1 + \beta^2) * (precision * recall)}{(\beta^2 * precision + recall)}$	Harmonic mean between precision and recall.

In the f-measure equation, β weights the importance of recall against precision. Because recall is the more important factor, a β value of 2 was used. A β of 2 has previously been used to evaluate extraction of problems from narrative text using NLP.[73]

The same measures were used to assess the performance of the problem list and to allow for comparison between the gold standard and problem list entries, with an emphasis on recall. Precision and f-measure were calculated, but these measures were less meaningful due to the likely validity of problem list false positives.

A 2-sample test of proportion was used to determine whether the evaluation metrics for problem list entries were significantly different from the evaluation metrics for nCode (Stata 11,

StataCorp LP, College Station TX). This test was also used to compare precision by Denny et al in a comparison of chart review, billing records review, and NLP for identification of colorectal cancer testing.[74]

Error analysis

An error analysis was performed to categorize the nature of the mismatches between the system and the gold standard list. False positive and false negative mismatches were examined and classified. Concept relations within the UMLS were used to determine whether two concepts were partial matches. If a qualifier was missing from a gold standard problem (i.e. *abscess* extracted instead of *epidural abscess*), this was also noted as a partial match.

RESULTS

The gold standard list contained 170 problems over 74 notes, or an average of 2.3 problems per consultation note. The gold standard lists contained a minimum of one problem and maximum of 7 problems per note. In total, there were 860 NLP extracted problems for 74 notes, or an average of 11.6 problems extracted per note. Discern nCode identified a minimum of 2 and maximum of 32 problems per note. The most common problems were HIV and MRSA infections.

Performance Evaluation

Discern nCode correctly represented 111 problems (Table 1). The average recall of Discern nCode per note was 65% and the precision was 14%, resulting in an F-measure of 37% (Table 3). The problem lists contained 166 problems; of these, 40 had been recorded within one week of the gold standard note. The recall of the entire problem list for the gold standard problems was 26%, higher than for the problem list entries that were made within a week of the note date (Table 3).

Table 1. Concepts identified by Discern nCode

	Gold Standard (+)	Gold Standard (-)
Discern nCode (+)	111	688
Discern nCode (-)	61	N/A

Table 2. Concepts identified in problem list

	Gold Standard (+)	Gold Standard (-)
Problem list (+)	17	23
Problem list (-)	153	N/A

Table 3. Evaluation metrics

	Discern nCode	Problem list
Recall	0.65	0.10*
Precision	0.14	0.43*
F-measure	0.37	0.12*

*$p < .05$

Natural language processing error analysis

False negatives

The system's recall is the most important aspect in this evaluation, and there were several major reasons for problems being identified as false negatives, with some problems falling into multiple categories (Table 4). Of the false negatives, 30 were abbreviations.

Table 4. Types of false negatives

Category	**# Missed (%)**
Had partial match	16 (26%)
Had no match	45 (14%)
Misspelled	5
Term lacked concept code	7

Partial matches

A quarter of the gold standard problems partially matched an NLP extracted problem. The majority of these gold standard problems would have required the formation of compositional expressions in order for Discern nCode to capture the full concept. For example, the gold standard list contained "Campylobacter colitis," and the system detected *colitis* and *genus campylobacter*, but not together. The same was true of "C diff colitis," "MRSA abscess," "MRSA cellulitis," and "MSSA bacteremia." It was also common that a qualifier was not combined with a condition. In the case of "viral myelitis," only *myelitis* was identified; for "subacute bacterial meningitis," only *bacterial meningitis* was extracted. In the majority of these cases, the system extracted a problem that was more general than the gold standard concept.

No match

For nearly half the false negatives, there were no partial matches identified. It is unclear why some of these concepts were missed, as many had corresponding concept codes identified in SNOMED-CT; notably, *hepatitis C* was never identified, and was missed five times. Other

missed concepts failed to be identified due to misspellings, potential ambiguity in phrasing, and lack of a clear concept code. The majority of these missed concepts were present in the note in abbreviated form (i.e. “klebs pneumo,” “HCV,” “PsA,” “SSTI,” “NSTI”*).*

Clinical domain challenges

There were five spelling or abbreviation errors which may have resulted in the lack of detection of these concepts. Misspelled problems included “tinia pedis” (*tinea pedis*), “W. bancroftii” (*W. bancrofti*), and “Acinetobacer baumanii” (*Acinetobacter baumanii*). There were also several concepts which may have been missed due to ambiguity or abbreviation. Although it was not always apparent why a problem was not extracted, they were sometimes presented in a way that may have excluded them from being obvious problems. The following examples demonstrate this: 1) “This would be a classic case of S. pneumoniae,” 2) “I was asked to see the patient […] for evaluation of bacteremia,” 3) “Cx w 2 types x E Coli.”

These show that the indication of a problem’s existence may be lost due to phrasing. “Would” is often used to indicate conditionality, but is not used this way in example #1. Example #2 implies that the patient has bacteremia, and the rest of the note confirms this, but “evaluation of bacteremia” is not conclusive and may not have triggered the extraction of the concept. Example #3 demonstrates a heavily abbreviated sentence structure which makes it difficult to tell that the culture was growing E. coli.

Vocabulary limitations

There were several false negatives for which there was no concept code available in SNOMED-CT. The most common of these included hardware-associated infections. Though there are several codes in SNOMED-CT related to this (e.g., *Infection and inflammatory reaction due to other internal orthopaedic prosthetic devices, implants and grafts*), “hardware-associated

infxn," and several variants of this, were not identified. "Viral respiratory illness" was written in the note but not identified; the more accurate "viral respiratory infection" has a code in SNOMED CT and may have been extracted. Another concept not extracted was "vascular infection," which did not clearly match with a concept code in SNOMED CT. Problems that were shunt-related, catheter-related or catheter-associated were only partially extracted, despite the existence of these concepts in SNOMED CT.

False positives

False positives within the notes fell into three general categories: problems that were exact concept matches but were not significant to the problem list, mismatched concepts, or falsely identified concepts.

Table 5. Types of false positives

Category	**# FPs (%)**
Exact match	480 (70%)
Judged not appropriate	259
Past medical history	221
Partial match	80 (11%)
More general concept	63
More specific concept	17
No match	128 (19%)
Word sense disambiguation	29
Unclear why extracted	26
Contextual	22
Uncertainty	19
Related concept	17
Negation	12
Spelling	3

For the majority of false positives, an exact match could be located within the note. These cases were predominantly false positives because they were judged not appropriate to the problem list

due to their absence from the gold standard, either as non-infectious diseases, or because they were not the preferred way of expressing a problem.

Many extracted false positives were concept matches, but were listed under the Past Medical History section of the note and therefore not judged to be ongoing problems by the gold standard. Some past medical problems were also mentioned in other areas of the note, and were either stated explicitly as being in the past or temporality was implied. These might be phrased as "h/o tinea dermatitis," "medical history significant for type 2 diabetes," "hx of Campylobacter colitis," "left ankle fracture 1986," "he had LOC for about 10 seconds."

There were numerous instances of Discern nCode extracting a problem which a human likely would not have, given the context. These reasons are described below:

Word sense

Many abbreviations were misidentified. The most common of these overloaded abbreviations included *RA* meaning *room air* and not *rheumatoid arthritis*, *MS* meaning *medical student* and not *multiple sclerosis*, *OI* meaning *opportunistic infection* and not *osteogenesis imperfecta*, *HD* meaning hemodialysis and not *Hodgkin's lymphoma*, and *HCM* meaning *health care management* and not *hypertrophic cardiomyopathy*.

Rent deposit was extracted as "deposition." Burn was extracted when it referred to the Burn/Plastics Clinic, or in the phrases "burned through multiple regimens" or "has burned many bridges with treatment facilities." Patient is "anxious to leave the hospital" resulted in extraction of *anxiety*. "Multiple arrests/short stints in jail" resulted in extraction of *asystole*. "Would consider stopping it cold turkey" resulted in extraction of *common cold*.

Contextual

Other people: “family etoh abuse,” “mother died of MVC.”

Patient-reported problems: “Endorses PTSD over schizoaffective disorder, with very poor insight”; the implication is that the patient does not have PTSD, but PTSD was extracted.

Not a problem: “Received influenza vaccine” was extracted as *influenza*, patient is “anxious about the price of antibiotics” resulted in extraction of *anxiety*. “Mood component of bipolar” resulted in extraction of *bipolar disorder*.

Uncertainty

Potential current problems: “Most likely staph,” “likely due to primary hyperparathyroidism,” “possible GC/CT,” “rule out osteopenia/osteoporosis,” “include flu treatment until PCR returns,” “whether chronic lymphedema vs CHF,” “he could be presenting with,” “concerning for a demyelinating process such as MS,” “may be due to peroneal tendon injury.”

Potential future problems: “This organism can commonly lead to intracerebral abscesses,” “Follow-up HIV, Hep B/C panels,” “Additional complications include cerebral edema.”

Related concept

Some of the false positives did not qualify as partial matches but were somewhat related because Discern nCode picked up on part of the concept (for example, extracted *systolic heart failure* when the note mentioned “decompensated heart failure,” *deficiency of testosterone biosynthesis* when the note mentioned “low testosterone,” and *depletion* when the note mentioned “chronic diffuse cerebral volume loss”).

Negation

"She is MRSA nasal swab –ve," "there has never been any evidence of active TB," "fungal NG," "no periosteal reaction or radiographic evidence of osteomyelitis," "influenza negative", "denies DM, HTN, cancer."

DISCUSSION

If used after the writing of a consultation note, NLP could speed the addition of problem list entries and increase the overall sensitivity of the problem list. This study demonstrates that Discern nCode was more sensitive for infectious disease problems from consultation notes than the post-visit problem list, though it had a lower precision.

The sensitivity of the problem list for gold standard problems was not dissimilar from other studies. Meystre and Haug found that the sensitivity of the problem list in a subset of patients at Intermountain Healthcare was 4% prior to intervention, demonstrating that the problem list was hardly used for documentation.[31] In this study, the average sensitivity of post-visit problem list entries was 10%, and the sensitivity of the entire problem list for gold standard entries was 26%. Although the low sensitivity of the problem list is mainly due to a lack of documentation, it is also a result of our decision to only consider exact matches, and not partial matches (e.g., *Osteomyelitis of vertebra* and *Osteomyelitis*). Additionally, alternative terms were sometimes chosen to capture a problem (for example, the problem list might contain the name of the organism causing the infection (e.g. *Staphylococcus aureus*) while the gold standard might have called out the syndrome (e.g. bacteremia) that was due to *Staphylococcus aureus*). There are many clinically similar concepts that are conceptually different within the UMLS. This can make it difficult to assess the sensitivity of the problem list, though the use of a pick list helps to reduce these variations. For example, *wound*, *abscess*, *cellulitis*, and *soft tissue infection* may all be used when referring to the same problem in a single patient, and it would be redundant to add all of them to the problem list.

The performance of the natural language processing system in this study was lower than in comparable studies, though it performed better than our measures reflect, especially due to the strict definition of a true positive; many of the partial matches may have been viable options for

problem entry. The evaluation studies of the MCVS, described earlier, found sensitivities ranging from 87-100%.[42,66] However, in both cases, the tasks (e.g., identification of quality indicators or pneumonia) involved a more well-defined set of potential concepts to extract, while identification of infectious disease problems involves a much greater range of possibilities as well as variation in how a problem can be expressed. The good performance of the system used by Meystre and Haug[36] is likely also due to the limited set of problems; 80 problems were considered and thus were prone to less variation than the gold standard problems created for this study. By contrast, like Solti et al,[60] I did not limit annotators to a predefined set of problems in order to more accurately reflect clinical practice. This means that problems could be entered at varying levels of granularity (from *cellulitis* to *hardware-associated MRSA cellulitis of the lower left extremity*), resulting in many partial matches. Solti et al found a higher system performance (88% sensitivity, 66% precision) but their evaluation methods deviated in some ways, such as qualifiers not factoring into the assessment of true positives.

Because the gold standard in this study was limited to infectious disease diagnoses and Discern nCode was not modified to this specific diagnosis classification, most of the false positives were due to the extraction by Discern nCode of non-infectious disease problems, and were false positives only in that they extended beyond the arbitrary boundaries used for this study. Modification to screen out non-infectious diseases would likely greatly improve system precision for this particular study, but this was not performed due to time constraints and the greater emphasis on recall. In addition, depending on how the filtering is done, this could result in an increase in false negatives; some findings and disorders such as *dog bite* might be clinically relevant to infectious disease without being classified as such within a controlled vocabulary. The system also did not screen out past medical problems, which accounted for a quarter of the false positives. It is unclear whether this modification would be useful; while it would increase

the precision of the problem list to reflect *current* problems, older diagnoses may not be resolved, and are often relevant in the care of current, active illnesses.

The analysis of false negatives demonstrated the significance of compositionality. Because it is important to prevent the problem list from being unnecessarily cluttered and to ensure that it best represents a patient's problems, gold standard problems commonly involved several concepts. Many of the missed concepts had partial concept matches identified. Improving the ability of the system to provide compositional expressions would reduce the number of false negatives. Though many missed problems would have involved stringing together multiple concepts, a minimal number of them were composed of terms which lacked corresponding concept codes. A 2010 survey of SNOMED CT users found that only 58% of users were satisfied or very satisfied with SNOMED CT's content coverage,[75] but SNOMED CT has better content coverage and is significantly more complete than other coding schemes,[14,26,76–78] and this was reflected by the results of this study.

William Long extracted diseases from discharge summaries and performed an analysis of missed concepts and false positives. In line with our findings, he found that most false positives were insignificant findings, historical, or ambiguous (i.e. possible, potential). He found that false negatives occurred because inference was needed, phrases were missing words, or phrases contained extra words which obscured the concept.[59] By contrast, Meystre and Haug found numerous false positives were due to negation detection errors and failure to incorporate contextual information.[36] In our study, negation errors did not result in many false positives.

Study limitations

This study had several limitations; according to the best practices for NLP evaluation studies described by Hripscak and Friedman, using even three experts may not be enough to establish a

reliable gold standard.[79] Given how difficult it is to obtain high inter-rater agreement when establishing a problem list,[71] it would have been ideal to use at least three annotators for each note and to have had annotation performed by infectious disease domain experts, especially since I was selecting charts in which problem list entries were or would have been made by infectious diseases physicians. These are common constraints; a 2010 systematic review of automated coding and classification systems in healthcare found that 43% of studies used a gold standard created by two or more independent reviewers, while the majority used one human reviewer, and the minority used a "trained standard" wherein an expert reviewer performs the majority of the work and training is provided to ensure consistency and performance.[80] Using explicit review methods for the annotations, and limiting gold standard entries to problems from the ID pick list, would have also increased inter-rater agreement. In addition, this study was performed within the context of infectious diseases and its results may not extend to other domains, but it still demonstrates some potential areas for improvement of system performance.

System recommendations

I recommend several ways of improving system performance, including allowing the combining of organism names (e.g. *Staphylococcus aureus*) with disease syndrome names (e.g. bacteremia) to increase recall. Recall could be enhanced by modifying rules within Discern nCode to handle abbreviations and variations that led to missed diagnoses. Use of novel assertion classification approaches as described by Bejan et al[81] could increase the precision and recall of nCode by improving how it handles assertion categories, particularly those that are conditional, hypothetical, or possible. Word sense errors were frequent, and incorporating word sense disambiguation methods could improve system performance.

Many extracted false positives may be useful, with some offering differing levels of granularity for problems within the note. Screening out problems within the Past Medical History section

would reduce the number of false positive diagnoses per note. However, this would hinder the care of patients, as past problems are often highly relevant to the management of current/active problems. Given that the average number of extracted problems per note was 11.6, and that many false negatives were potentially applicable, the suggestions offered by nCode could be rapidly screened to facilitate problem entry. Thus, future efforts should focus on increasing the sensitivity (recall) of Discern nCode, while reduction of false positives is a lower priority given the ease of screening.

Conclusion

Though variation in problem lists is common and the expression of many problem variants should be supported, future research should also explore ways to make problem lists, and problem list entry guidelines, more consistent. Deciding what should be on the problem list, and who should put it there, is an integral component of establishing a complete and accurate list. However, making problem entry easier and more intuitive is also critical. Discern nCode outperformed current problem documentation efforts. If used for automatic prompting, it would likely increase problem list completeness, though efforts should focus on improving system recall.

ACKNOWLEDGEMENTS

I wish to thank Drs. Thomas Payne, Robert Harrington and Meliha Yetisgen-Yildiz for their guidance, feedback, and assistance with the preparation of this manuscript. I would also like to thank Drs. Andrew White and Daniel Martin for assisting with data collection as annotators.

BIBLIOGRAPHY

1. Weed, L. L. Medical records that guide and teach. *The New England journal of medicine* **278**, 593–600 (1968).

2. Dick, R., Steen, E. & Detmer, D. *The Computer-based Patient Record: An Essential Technology for Health Care*. (National Academy Press: Washington, DC, 1997).

3. *Comprehensive Accreditation Manual for Hospitals*. (Oakbrook Terrace, IL, 1996).at <http://www.jcaho.org>

4. Health information technology: initial set of standards, implementation specifications, and certification criteria for electronic health record technology. Interim final rule. *Federal register* **75**, 2013–47 (2010).

5. Hartung, D. M., Hunt, J., Siemienczuk, J., Miller, H. & Touchette, D. R. Clinical implications of an accurate problem list on heart failure treatment. *Journal of general internal medicine* **20**, 143–7 (2005).

6. Benson, D. S., Van Osdol, W. & Townes, P. Quality ambulatory care: the role of the diagnostic and medication summary lists. *QRB* **14**, 192–7 (1988).

7. Margolis, C. Z., Newborn, J. L. & Sheehan, T. J. Effect of the problem oriented record system on care in a pediatric clinic. *Pediatric research* **13**, 1047–51 (1979).

8. Simborg, D. W., Starfield, B. H., Horn, S. D. & Yourtee, S. A. Information factors affecting problem follow-up in ambulatory care. *Medical care* **14**, 848–56 (1976).

9. Poon, E. G. *et al.* Relationship between use of electronic health record features and health care quality: results of a statewide survey. *Medical care* **48**, 203–9 (2010).

10. Bayegan, E. & Tu, S. The helpful patient record system: problem oriented and knowledge based. *Proceedings of the AMIA Annual Symposium* 36–40 (2002).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2244287&tool=pmcentrez&rendertype=abstract>

11. Galanter, W. L., Hier, D. B., Jao, C. & Sarne, D. Computerized physician order entry of medications and clinical decision support can improve problem list documentation compliance. *International journal of medical informatics* **79**, 332–8 (2010).

12. Sarkar, U. *et al.* SynopSIS: integrating physician sign-out with the electronic medical record. *Journal of hospital medicine : an official publication of the Society of Hospital Medicine* **2**, 336–42 (2007).

13. Zelingher, J. *et al.* Categorization of free-text problem lists: an effective method of capturing clinical data. *Proceedings of the Annual Symposium on Computer Application in Medical Care* 416–20 (1995).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2579126&tool=pmcentrez&rendertype=abstract>

14. Campbell, J. R. & Payne, T. H. A comparison of four schemes for codification of problem lists. *Proceedings of the Annual Symposium on Computer Application in Medical Care* 201–5 (1994).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2247911&tool=pmcentrez&rendertype=abstract>

15. Johnston, M. E., Langton, K. B., Haynes, R. B. & Mathieu, A. Effects of computer-based clinical decision support systems on clinician performance and patient outcome. A critical appraisal of research. *Annals of internal medicine* **120**, 135–42 (1994).

16. Kaushal, R., Shojania, K. G. & Bates, D. W. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. *Archives of internal medicine* **163**, 1409–16 (2003).

17. Wright, A. *et al.* Ability to generate patient registries among practices with and without electronic health records. *Journal of medical Internet research* **11**, e31 (2009).

18. Gundersen, M. L. *et al.* Development and evaluation of a computerized admission diagnoses encoding system. *Computers and biomedical research, an international journal* **29**, 351–72 (1996).

19. Haug, P. J. *et al.* A natural language parsing system for encoding admitting diagnoses. *Proceedings of the AMIA Annual Fall Symposium* 814–8 (1997).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2233343&tool=pmcentrez&rendertype=abstract>

20. Wright, A. *et al.* A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. *Journal of the American Medical Informatics Association : JAMIA* **18**, 859–67 (2011).

21. Jao, C. S., Hier, D. B. & Galanter, W. L. Using clinical decision support to maintain medication and problem lists A pilot study to yield higher patient safety. *2008 IEEE International Conference on Systems, Man and Cybernetics* 739–743 (2008).doi:10.1109/ICSMC.2008.4811366

22. Hripcsak, G. *et al.* Unlocking clinical data from narrative reports: a study of natural language processing. *Annals of internal medicine* **122**, 681–8 (1995).

23. Meddings, J., Saint, S. & McMahon, L. F. Hospital-acquired catheter-associated urinary tract infection: documentation and coding issues may reduce financial impact of Medicare's new payment policy. *Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America* **31**, 627–33 (2010).

24. Classen, D. C. *et al.* "Global trigger tool" shows that adverse events in hospitals may be ten times greater than previously measured. *Health affairs (Project Hope)* **30**, 581–9 (2011).

25. Bates, D. W. *et al.* Detecting adverse events using information technology. *Journal of the American Medical Informatics Association : JAMIA* **10**, 115–28 (2003).

26. Elkin, P. L. *et al.* A randomized controlled trial of the accuracy of clinical record retrieval using SNOMED-RT as compared with ICD9-CM. *Proceedings of the AMIA Annual Symposium* 159–63 (2001).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2243271&tool=pmcentrez&rendertype=abstract>

27. Rosof, B. The importance of accurate data in quality-of-care measurement. *Annals of internal medicine* **157**, 379–80 (2012).

28. Szeto, H. C., Coleman, R. K., Gholami, P., Hoffman, B. B. & Goldstein, M. K. Accuracy of computerized outpatient diagnoses in a Veterans Affairs general medicine clinic. *The American journal of managed care* **8**, 37–43 (2002).

29. Tang, P. C., LaRosa, M. P. & Gorden, S. M. Use of computer-based records, completeness of documentation, and appropriateness of documented clinical decisions. *Journal of the American Medical Informatics Association : JAMIA* **6**, 245–51 (1999).

30. Wright, A., Maloney, F. L. & Feblowitz, J. C. Clinician attitudes toward and use of electronic problem lists: a thematic analysis. *BMC medical informatics and decision making* **11**, 36 (2011).

31. Meystre, S. M. & Haug, P. J. Randomized controlled trial of an automated problem list with improved sensitivity. *International journal of medical informatics* **77**, 602–12 (2008).

32. Hier, D. B. Unpublished Audit of Medical Records at the University of Illinois Hospital. (2002).

33. Campbell, J. R. Strategies for problem list implementation in a complex clinical enterprise. *Proceedings of the AMIA Annual Symposium* 285–9 (1998).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2232208&tool=pmcentrez&rendertype=abstract>

34. Holmes, C. The problem list beyond meaningful use. Part I: The problems with problem lists. *Journal of AHIMA / American Health Information Management Association* **82**, 30–3; quiz 34 (2011).

35. Kaplan, D. M. Clear writing, clear thinking and the disappearing art of the problem list. *Journal of hospital medicine : an official publication of the Society of Hospital Medicine* **2**, 199–202 (2007).

36. Meystre, S. & Haug, P. J. Automation of a problem list using natural language processing. *BMC medical informatics and decision making* **5**, 30 (2005).

37. Meystre, S. & Haug, P. J. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *Journal of biomedical informatics* **39**, 589–99 (2006).

38. Meystre, S. & Haug, P. Improving the sensitivity of the problem list in an intensive care unit by using natural language processing. *Proceedings of the AMIA Annual Symposium*

554–8 (2006).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1839473&tool=pmcentrez&rendertype=abstract>

39. Wang, X., Hripcsak, G., Markatou, M. & Friedman, C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association : JAMIA* **16**, 328–37 (2009).

40. Xu, H. *et al.* MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association : JAMIA* **17**, 19–24 (2010).

41. Gold, S., Elhadad, N., Zhu, X., Cimino, J. J. & Hripcsak, G. Extracting structured medication event information from discharge summaries. *Proceedings of the AMIA Annual Symposium* 237–41 (2008).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2655993&tool=pmcentrez&rendertype=abstract>

42. Elkin, P. L. *et al.* NLP-based identification of pneumonia cases from free-text radiological reports. *Proceedings of the AMIA Annual Symposium* 172–6 (2008).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2656026&tool=pmcentrez&rendertype=abstract>

43. Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J. & Johnson, S. B. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association : JAMIA* **1**, 161–74 (1994).

44. Yetisgen-Yildiz, M. & Glavan, B. Identifying patients with pneumonia from free-text intensive care unit reports. *The International Conference on Machine Learning* (2011).at <http://staff.washington.edu/melihay/publications/ICML_2011.pdf>

45. Deleger, L. *et al.* Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *Journal of the American Medical Informatics Association : JAMIA* **20**, 84–94 (2013).

46. Matheny, M. E. *et al.* Detection of infectious symptoms from VA emergency department and primary care clinical documentation. *International journal of medical informatics* **81**, 143–56 (2012).

47. Ruch, P. & Gaudinat, A. Comparing corpora and lexical ambiguity. *Proceedings of the workshop on Comparing corpora* - **9**, 14–19 (2000).

48. Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C. & Hurdle, J. F. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics* 128–44 (2008).at <http://www.ncbi.nlm.nih.gov/pubmed/18660887>

49. Kleinsorge, R. & Willis, J. Unified Medical Language System (UMLS) Basics. *National Library of Medicine* 1–157at <http://www.nlm.nih.gov/research/umls/pdf/UMLS_Basics.pdf>

50. Liu, H., Lussier, Y. A. & Friedman, C. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. *Journal of biomedical informatics* **34**, 249–61 (2001).

51. Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F. & Buchanan, B. G. Evaluation of negation phrases in narrative clinical reports. *Proceedings of the AMIA Annual Symposium* 105–9 (2001).at </pmc/articles/PMC2243578/?report=abstract>

52. Liu, F., Weng, C. & Yu, H. *Clinical Research Informatics*. *Clinical Research Informatics* 293–310 (Springer London: London, 2012).doi:10.1007/978-1-84882-448-5

53. Chapman, W. W. Natural Language Processing for Biosurveillance. *Handbook of biosurveillance* 255–71 (2011).

54. D'Avolio, L., Demner-Fushman, D. & Chapman, W. An Introduction to Clinical Natural Language Processing. *AMIA Annual Symposium* (2011).at <http://idash.ucsd.edu/sites/default/files/nlp-media/AMIA-NLPPart1-10182011.pdf>

55. Chute, C. G., Yang, Y. & Buntrock, J. An evaluation of computer assisted clinical classification algorithms. *Proceedings of the Annual Symposium on Computer Application in Medical Care* 162–6 (1994).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2247905&tool=pmcentrez&rendertype=abstract>

56. Pakhomov, S. V. S., Buntrock, J. D. & Chute, C. G. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association : JAMIA* **13**, 516–25 (2006).

57. Haug, P. *et al.* A natural language understanding system combining syntactic and semantic techniques. *Proceedings of the Annual Symposium on Computer Application in Medical Care* 247–51 (1994).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2247803&tool=pmcentrez&rendertype=abstract>

58. Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F. & Buchanan, B. G. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics* **34**, 301–10 (2001).

59. Long, W. Lessons extracting diseases from discharge summaries. *Proceedings of the AMIA Annual Symposium* 478–82 (2007).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2655845&tool=pmcentrez&rendertype=abstract>

60. Solti, I. *et al.* Building an automated problem list based on natural language processing: lessons learned in the early phase of development. *Proceedings of the AMIA Annual Symposium* 687–91 (2008).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2655946&tool=pmcentrez&rendertype=abstract>

61. Elkin, P. L. *et al.* Standardized problem list generation, utilizing the Mayo canonical vocabulary embedded within the Unified Medical Language System. *Proceedings of the AMIA Annual Fall Symposium* 500–4 (1997).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2233586&tool=pmcentrez&rendertype=abstract>

62. Elkin, P. L. *et al.* The role of compositionality in standardized problem list generation. *Studies in health technology and informatics* **52 Pt 1**, 660–4 (1998).

63. Elkin, P. L. *et al.* Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. *Mayo Clinic proceedings. Mayo Clinic* **81**, 741–8 (2006).

64. Elkin, P. L. *et al.* A controlled trial of automated classification of negation from clinical notes. *BMC medical informatics and decision making* **5**, 13 (2005).

65. Brown, S. H. *et al.* eQuality: electronic quality assessment from narrative clinical reports. *Mayo Clinic proceedings. Mayo Clinic* **81**, 1472–81 (2006).

66. Brown, S. H., Elkin, P. L., Rosenbloom, S. T., Fielstein, E. & Speroff, T. eQuality for all: Extending automated quality measurement of free text clinical narratives. *Proceedings of the AMIA Annual Symposium* 71–5 (2008).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2656015&tool=pmcentrez&rendertype=abstract>

67. Welton, N. J. The University of Washington electronic medical record experience. *Journal of the Medical Library Association : JMLA* **98**, 217–9 (2010).

68. Van Eaton, E. G., Horvath, K. D., Lober, W. B., Rossini, A. J. & Pellegrini, C. A. A randomized, controlled trial evaluating the impact of a computerized rounding and sign-out system on continuity of care and resident work hours. *Journal of the American College of Surgeons* **200**, 538–45 (2005).

69. US Department of Health and Human Services Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. (2010).at <http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html>

70. Hripcsak, G. & Rothschild, A. S. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association : JAMIA* **12**, 296–8 (2005).

71. Rothschild, A. S., Lehmann, H. P. & Hripcsak, G. Inter-rater agreement in physician-coded problem lists. *Proceedings of the AMIA Annual Symposium* 644–8 (2005).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1560827&tool=pmcentrez&rendertype=abstract>

72. Resnik, P. & Lin, J. 11 Evaluation of NLP Systems. *The Handbook of Computational Linguistics and Natural Language Processing* 271–295 (2010).

73. Meystre, S. M. & Haug, P. J. Comparing natural language processing tools to extract medical problems from narrative text. *Proceedings of the AMIA Annual Symposium* 525–9 (2005).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1560561&tool=pmcentrez&rendertype=abstract>

74. Denny, J. C. *et al.* Natural language processing improves identification of colorectal cancer testing in the electronic medical record. *Medical decision making : an international journal of the Society for Medical Decision Making* **32**, 188–97 (2012).

75. Elhanan, G., Perl, Y. & Geller, J. A Survey of Direct Users and Uses of SNOMED CT: 2010 Status. *Proceedings of the AMIA Annual Symposium* **2010**, 207–11 (2010).

76. Campbell, J. R. *et al.* Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. CPRI Work Group on Codes and Structures. *Journal of the American Medical Informatics Association : JAMIA* **4**, 238–51 (1997).

77. Chute, C. G., Cohn, S. P., Campbell, K. E., Oliver, D. E. & Campbell, J. R. The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. *Journal of the American Medical Informatics Association : JAMIA* **3**, 224–33 (1996).

78. Wasserman, H. & Wang, J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. *Proceedings of the AMIA Annual Symposium* 699–703 (2003).at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1479961&tool=pmcentrez&rendertype=abstract>

79. Friedman, C. & Hripcsak, G. Evaluating natural language processors in the clinical domain. *Methods of information in medicine* **37**, 334–44 (1998).

80. Stanfill, M. H., Williams, M., Fenton, S. H., Jenders, R. A. & Hersh, W. R. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association : JAMIA* **17**, 646–51 (2010).

81. Bejan, C. A., Vanderwende, L., Xia, F. & Yetisgen-Yildiz, M. Assertion modeling and its role in clinical phenotype identification. *Journal of biomedical informatics* **46**, 68–74 (2013).