Concept-Value Pair Extraction from Semi-Structured Clinical Reports: A Case Study Using Echocardiogram Reports

by
Jeanhee Chung

Submitted to the Health Sciences and Technology Division and the
Division of Medical Informatics
In Partial Fulfillment of the Requirements for the
Degrees of Master of Science in Medical Informatics
At the Massachusetts Institute of Technology
May 20, 2004
[June 2004]

Signature of Author..
Health Sciences and Technology
May 20, 2004

Certified by..
G. Octo Barnett M.D.
Professor of Medicine, Harvard Medical School
Thesis Supervisor

Certified by..
Henry Chueh M.D.
Assistant Professor of Medicine, Harvard Medical School
Thesis Supervisor

Certified by..
Shawn N. Murphy M.D., Ph.D.
Instructor of Medicine, Harvard Medical School
Director, Research Patient Data Registry, Massachusetts General Hospital
Research Supervisor

Accepted by..
Martha L. Gray PhD
Edward Hood Taplin Professor of Medical and Electrical Engineering
Co-director, Harvard - MIT Division of Health Sciences and Technology

Concept-Value Pair Extraction from Semi-structured Clinical Reports: A Case Study Using Echocardiogram Reports

by
Jeanhee Chung

Submitted to the Health Sciences and Technology Division and the
Division of Medical Informatics

May 20, 2004

In Partial Fulfillment of the Requirements for the
Degrees of Master of Science in Medical Informatics
At the Massachusetts Institute of Technology

ABSTRACT

The task of gathering detailed patient information from narrative clinical text presents a significant barrier to clinical research. A prototype information extraction system was developed to extract pre-specified findings from narrative echocardiogram reports. The system which uses a Unified Medical Language System compatible architecture is very simple and takes advantage of canonical language use patterns to identify sentence templates with which concepts and their values can be identified. The data extracted from this system will be used to enrich an existing database used by clinical researchers in a large university healthcare system to identify potential research candidates fulfilling clinical inclusion criteria. The system was developed and evaluated using ten pre-determined clinical concepts. Concept-value pairs extracted by the system related to these ten conditions were compared with findings extracted manually by the author. The system was able to recall 78% of the relevant findings (CI, 76% to 80%), with a precision of 99% (CI, 98%-99%). Because data acquired from the system will ultimately be used in document and patient retrieval, preliminary analysis was done to evaluate document retrieval effectiveness. Median recall across the ten conditions was 36% (range, 0% to 93%). The system retrieved no documents for two of the ten conditions; median precision for the remaining eight conditions was 100% (range, 92% to 100%).

Keywords: information extraction; natural language processing

Thesis Supervisor: G. Octo Barnett, M.D.
Title: Professor of Medicine, Harvard Medical School

Acknowledgements

The author would like to thank the following for their generous contributions of time, knowledge, and support:

- My research and thesis supervisors, Shawn N. Murphy, G. Octo Barnett, and Henry Chueh, without whose support and patience this work would not have been possible;
- The National Library of Medicine and the training grant for medical informatics, for its outstanding support;
- All my colleagues at the Laboratory of Computer Science who each helped tremendously in their own ways;
- And especially all my family and friends who provided continuous encouragement and support.

Table of Contents

Introduction

The desire to use automated systems to enhance clinical care and facilitate clinical research has become pervasive [1, 2]. To function properly, these systems require accurate, complete data that can be systematically accessed. In order to increase the value of electronic record information, information must be represented in a way suitable for automated retrieval and processing [3]. To be accurate, automated systems require *coded data*: the concepts must come from a well-defined, finite vocabulary, and the relations among the concepts must be expressed in an unambiguous, formal structure.

Clinical records, however, frequently contain a mixture of highly structured or coded data along with loosely structured narratives [4]. The coded information can be easily retrieved, but clinical data in the narrative form is not accessible except by manual review. While personnel can be trained to read and manually structure reports [5], few institutions are willing to invest in manual coding (other than for billing purposes). Ultimately, collecting highly differentiated and specific clinical data from electronic patient records continues to be a major obstacle for taking full advantage of clinical information systems in health care [6].

One strategy to deal with the dearth of useable narrative information has been to prevent more of it from being produced. In this scenario, authors of clinical reports (such as radiologists who interpret chest x-rays) are trained to enter only coded clinical data. In some cases, this data is converted into some kind of narrative to mimic the "natural text" to which the consumers of these reports are accustomed. This method has met with success in certain domains, such as laboratory result reporting, and to some extent in radiology and cardiology. However, this method of entering patient information is considered too restrictive for most clinical purposes and the computer-generated narrative is frequently altered to fit the author's needs. In fact, it is not uncommon to find the coded sections left blank and the narrative section filled instead. In these cases, the narrative contains important clinical data that cannot be found elsewhere in the document.

Another strategy is to permit narrative input which can then be processed to structure and codify the information it contains. However, the descriptive flexibility that writing narrative allows, in turn makes processing this narrative challenging. Sentences that are easy for a person to understand are difficult for a computer to sort out. While this solution necessitates fewer workflow changes while automating access to narrative data, development of such systems requires highly specialized linguistic and computational expertise and support.

A third intermediate alternative is to focus on extracting only certain data elements residing in the narrative text—rather than trying to achieve full natural language understanding. Information extraction systems automatically extract unstructured or partially structured information from machine-readable files. These automated systems can be used to pull out pertinent data from the clinical narrative without requiring full language analysis. Although these systems can still be quite complex to develop, the narrowed scope of the problem permits simpler solutions that can achieve relatively good performance.

This work describes the implementation and evaluation of an information extraction system in which concept-based sentence templates are discovered from echocardiogram reports. These templates are then used to extract pre-specified clinical concepts and their associated values. While the intended target of this system will ultimately be all procedure-based reports, this paper demonstrates how this simple method can be applied to echocardiogram reports and yield patient-specific cardiac findings in a reliable way. In addition to evaluating the system's ability to extract information, secondary data is introduced demonstrating how it may aid document retrieval.

1 Background

1.1 The Medical Narrative [7]

The medical narrative has been described as a *sublanguage*. Sublanguages are the specialized and structured languages used in a scientific or technical domain. While a sublanguage shares many of the syntactic* forms of the parent language, it may add some of its own, or it may use the parent forms in different ways. For example, even though these sentences are grammatically correct in English, only one of them makes sense in biology: "Ions enter the cell" vs. "The cell enters ions".

Medical sublanguage grammar has remained stable over a range of clinical areas and the narrower vocabulary of these domains has required little modification. Medical narratives are found to contain only a limited number of types of statements, seen as patterns of sublanguage word classes (e.g. [DISEASE] or [FINDING]) with underlying grammatical relations. One of the first medical language processing endeavors, the Linguistic String Project, used *information formats* that served as templates for particular types of medical statements. It consisted of certain fixed fields, where each field corresponds to a particular type of medical information, e.g. [DISEASE] or [MEDICATION]. The conversion of information in documents from its free-text form to an 'informationally equivalent' structured form is possible because of the syntactic and semantic regularities present in the original documents. Some of the regularities are general for the English language, and others are specific to the biomedical and clinical domains.

Different types of medical narrative display differing degrees of language regularity. Reports from diagnostic procedures, such as radiology or cardiology reports, are simpler documents compared with reports detailing patient care, such as discharge summaries or out-patient clinic notes [8]. These latter notes represent a greater challenge, in part, because of the highly individualized experience described. They can range from brief utterances of a few acronyms, e.g. "48 yo c CAD p/w SOB admitted for CP, r/o'ed MI. D/C'ed home on ASA and Lopressor," to page-long detailed narratives chronicling the patient's medical, social and family histories.

In contrast, the terminology and its interrelations are much better defined in procedural reports. Statements in an echocardiogram report, for example, are by definition about the patient's heart. Similar to other areas of "objective" clinical reporting, e.g. the physical exam, the subject of the description is always implicitly the patient, descriptions generally refer only to the present and sentences are generally independent units of description [9].

* Syntax refers to the way words are arranged together (i.e., word order).

1.2 Medical Language Processing [10]

Natural language processing, along with computational linguistics and speech recognition and synthesis, comprise what is called *speech and language processing*. Speech and language processing encompasses everything from mundane applications such as word counting and automatic hyphenation to cutting edge applications such as automated question answering on the Web and speech recognition. It also includes *information retrieval* (finding out where textual resources reside), *information extraction* (extracting pertinent facts from those textual resources) and *inference* (drawing conclusions based on known facts). What distinguishes language-processing applications from other data processing systems is their use of language. Consider the **Microsoft Word** program that can be used to count the total number of words, lines and pages in a document. When used to count lines and pages, **Word** is an ordinary data processing application. However, when it is used to count the words in a file, it requires knowledge about what it means to be a word and thus it becomes a language processing system.

The knowledge of language needed to engage in complex language processing can be separated into six distinct categories:

1. Phonetics and phonology: the study of linguistic sounds;
2. Morphology: the study of meaningful components of words;
3. Syntax: the study of the structural relationships among words;
4. Semantics: the study of meaning;
5. Pragmatics: the study of how language is used to accomplish goals;
6. Discourse: the study of linguistic units larger than a single utterance.

For the purpose of the present discussion, it is sufficient to understand that the information extraction task on which we are about to embark, does not require the kind of complete language analysis assumed in the development of a comprehensive natural language processing system. It is hypothesized that procedural reports with their narrow terminology, little need for outside knowledge and predictable routine lend themselves to a comparatively shallower analysis.

1.2.1 Information Extraction [11, 12]

Information extraction (IE) is the process of extracting user-specified text from a set of documents. The extracted information includes entities, relations and most challenging, events. While it requires deeper analysis than key word searches, IE tasks are generally easier to implement than general-purpose natural language processing systems since complete syntactic characterization and language understanding are not necessary. The goal of IE is to capture structured information without sacrificing feasibility.

IE techniques attempt to identify semantic structure and other specific types of information from unrestricted text. These systems work particularly well in narratives in which the desired knowledge can be described by a relatively simple and fixed template, or frame, with slots that can be filled in with material from the text. Only a small part of the information in the text is relevant for filling in this frame; the rest can be ignored.

The most commonly used structure for IE systems are *cascaded finite-state transducers*, which separate processing into a series of steps. A finite-state automaton reads one word at a time in a sequence of words; each word transitions the automaton into a new state, based on its part of speech (or semantic category). Some states are designated as final, and a final state is reached when the sequence of words matches a valid pattern or a frame is complete. In a finite-state transducer, an output entity (e.g. a database entry) is constructed when final states are reached, e.g. a representation of the information in a phrase. In a cascaded finite-state transducer, there are different finite state transducers at different stages. Earlier stages will package a string of word elements into something that the next stage will view as a single element.

1.2.2 Information Extraction and Information Retrieval

Although information extraction systems and information retrieval systems are related, they represent two different processes. IR systems find collections of indexed documents pertaining to some criteria and present these to the user. The user would need to read the document in order to extract the appropriate information. Keyword-based document search, which is at the center of almost all document search tools today [13], is considered to be better at locating documents about topics rather than documents that report specific relationships [14].

If we are interested in repeated searches of a collection in a limited domain, it is possible to provide much more powerful search tools using IE techniques, which analyze the actual text. If the collection centers on a small number or relations or event types, it is possible to automatically extract these relations with links back to the original documents (or patients), which can then be retrieved.

1.2.3 Evaluation of Information Extraction Systems

The performance of both information extraction and information retrieval systems are assessed using similar metrics: precision and recall [10]. *Recall* or *sensitivity* is a measure of how much relevant information the system extracted from the text; it measures coverage of the system:

$$\text{Recall} = \frac{\text{\# of correct answers given by the system}}{\text{\# of possible correct answers in the text}}$$

Precision or *positive predictive value* is a measure of how much of the information extracted by the system is actually correct, and is also known as *accuracy*. Precision is defined as follows:

$$\text{Precision} = \frac{\text{\# of correct answers given by the system}}{\text{\# of answers given by the system}}$$

Note that recall and precision are antagonistic to one another since a conservative system that strives for perfection in terms of precision will invariably lower its recall score. Similarly, a system that strives for perfect coverage will get more things wrong, thus lowering its precision score.[Ω]

1.3 Brief Review of Information Extraction Efforts in the Medical Domain

In Message Understanding Conference[*] evaluations in the 1990s, names could be recognized at about 95% recall and precision— nearly human-level performance. Recognition of events and relations, however, plateaued at about 60% recall and precision [11].

The limited domain of biomedicine, however, offers an attractive opportunity for developers of information extraction systems. The combined advent of electronic clinical record keeping and rapid growth of biomedical literature have led to the eager anticipation that information extraction techniques will provide the key to unlock this data. Development in this area spans from simple keyword search tools to full natural language processing systems. This review is not intended to be comprehensive, but serves to give the reader an idea of the kinds of language applications being developed in clinical medicine and the range of results that has been produced by such systems.

In the clinical domain, specific types of reports have guided the development of extraction systems. Friedman et al. developed MedLEE[=] to extract findings from chest radiography reports; it has since been expanded to mammography reports, discharge summaries and pathology results [15-17]. One of the earlier studies involving MedLEE was a retrospective analysis in the detection of tuberculosis from chest radiography reports. In this setting, MedLEE demonstrated 89% agreement with a manually coded gold standard. Hripscak et al. later evaluated the performance of MedLEE with an automated decision-support system for six clinical conditions found in chest radiography reports and achieved a sensitivity of 81% with a specificity of 98%. 12 physicians (6 internists and 6 radiologists) were used to establish the reference standard.

[Ω] This situation has led to the use of a combined measure called the *F-measure* that balances recall and precision by using a parameter B. The F-measure is defined as follows:

$$F = \frac{(\mathrm{B}^2 + 1)(precision)(recall)}{\mathrm{B}^2 precision + recall}$$

When B is one, precision and recall are given equal weight. When B is greater than one, precision is favored, and when B is less than one, recall is favored. The more commonly used measures of recall and precision were used in the present study.

[*] Message Understanding Conferences were initiated by the Naval Ocean Systems Center to foster research in the automated analysis of military messages containing textual information. For each conference, participating groups are given sample messages and instructions on the type of information to be extracted. Their efforts have contributed greatly to promote and evaluate research in information extraction [13]

[=] Medical Language Extraction and Encoding System (MedLEE)

Interestingly, they observed that the physicians showed sensitivities and specificities that were more consistent with those of the processor in the study than in studies where the sensitivity was found to be higher than the specificity [18].

Multiple systems focusing on narrower domains have also been developed. Fiszman et al. developed SymText to extract relevant clinical information from ventilation/perfusion lung scan reports. The overall precision of their system was 88% with a recall of 92%. In the echocardiogram domain, Canfield et al. at the LDS hospital in Salt Lake City developed ECHODB to capture free text information in echocardiography reports using a frame data structure compatible with the Unified Medical Language System [19]. Haug et al. also at the LDS hospital, developed the SPRUS[<] system to extract findings and diagnostic interpretations from free-text chest x-ray reports using semantic information from a diagnostic expert system. They demonstrated a precision of 87% for radiographic findings and a precision of 95% for diagnostic interpretations [20].

Notably, all of the above systems (MedLEE, SymText, ECHODB, and SPRUS) are comprehensive systems that aim to capture all relevant information in their respective documents. Information extraction systems that do not rely on full parsing have also demonstrated promising results. Grishman et al. report on a system, Proteus-BIO, which provides a capability for searching documents on the Web about infectious disease outbreaks [14]. The basic structure of the Proteus extraction system uses a cascaded set of finite-state transducers described in 1.2.1. Sequences of words are ultimately mapped to 'event patterns' generating 'event structures' that are finally input into a database. Using a small test sample of 32 documents, their system demonstrated a precision of 79% with a recall of 41%. Lin et al. developed and evaluated a modular computer program that uses a user specified list of canonical phrases to automatically extract matched findings from dictated admission summaries. On a test set of 20 documents, they were able to show a recall of 92% with a precision of 96% [9].

Mikkelson and Aasly evaluated a template assisted manual semantic indexing of textual record notes. They found that models combining both simple string matching and semantic tagging performed better than separate models with respect to best mean recall (90%) and mean precision at high levels of recall (33%). Closer analysis of the study reveals that, only the tag-based system could yield a precision of 100% at a recall of 50% [21]. Unlike other systems that attempt to 'discover' templates from the text, this study created a pre-defined template that users subsequently had to follow. Brown et al. found that record retrieval utilizing the semantic characteristics of CLINICAL TERMS version 3 (Read Codes) performed significantly better than

[<] Special Purpose Radiology Understanding System (SPRUS) retrieval based on string matching with a mean recall of 94% vs. 61% and a mean precision of 99% vs. 82% in a database containing diabetic patient problem entries [22]. Barrows et al. compared a limited pattern matching system (Glaucoma Dedicated Parser- GDP) with MedLEE in the analysis of "notational" text visit notes in the detection of glaucoma diagnosis and progression. Recall and precision for the GDP over 7 parameters was 95-100% and 89-100%, respectively. Recall and precision for MedLEE over the same 7 parameters was 80-100% and 100%, respectively [23].

1.4 Lessons Learned

Retrieving information from narrative patient records is usually not an exact operation. As long as information is diversely represented, either by natural language or by structured data of variable consistency, requests for information return only parts that are relevant, accompanied by a varying amount of information that is irrelevant [21]. Extraction of information from such sources should pragmatically utilize any characteristics of the data that will increase retrieval effectiveness [21], at the same time recognizing that the nature of medical record keeping implies a high level of uncertainty and impaired reliability [24, 25]. String matching, concept pattern matching and pre-defined tagging methods have all been used successfully to locate information in narrative records. While these mechanisms may be insufficient to parse a complicated discharge summary, they have proven to be sufficient in information extraction tasks, particularly in the domain of procedural reporting.

1.4.1 Understanding Information Needs

In the absence of an ideal system which achieves both perfect recall and perfect precision, one needs to decide which parameter is more critical. This decision is application dependent and ultimately, depends on the needs of the individual user.

In clinical systems, such as automated alerting systems, the need for high sensitivity must be balanced with the need for high specificity. In such a system, high sensitivity would likely prevent the greatest number of adverse events. However, high specificity minimizes the false positive alerts that can potentially undermine user confidence in the automated system [1].

Researchers querying a clinical database of a million patients to identify a research cohort, however, have different needs. Depending on the cumulative prevalence of their research criteria in the database population, they do not necessarily need to identify *all* the patients who meet their inclusion criteria—they need to identify only some proportion of the group that *truly* meets their criteria. Precision becomes very important since time and resources will be expended for the subsequent chart review on the cases that the system does return.

Clearly, for both research and clinical data, the information must be accurate. While the consequences are arguably greater for incorrect information used directly in patient care, neither a clinician taking care of a patient nor a researcher who is looking through 10,000 patient charts want to falsely identify patients with any disease or disease attribute.

1.4.1.1 The Research Patient Data Registry (RPDR)

The Research Patient Data Registry (RPDR) is a clinical data warehouse developed for Partners Healthcare Inc. in Boston, which includes Massachusetts General Hospital (MGH), Brigham and Women's Hospital, Newton Wellesley Hospital and Faulkner Hospital [26]. The central role of this repository is to provide services to recruit patients for clinical research studies by providing electronic access to all inpatient and outpatient data. The kind of data identified as desirable for research was studied by Murphy et al. [27] by reviewing 16 years of COSTAR* research queries. Interestingly, they found that although researchers were most verbally interested in accessing data contained within narrative reports, most queries performed on the database did not involve narrative data at all. This discrepancy, however, may have been less a reflection of need than the relatively limited text processing capabilities of the query language used to access COSTAR.

The RPDR contains about 270 million patient-concept associations from 26 million patient encounters encompassing the 1.4 million patients in the Partners healthcare system. Clinical researchers use a visual query tool to identify cohorts of patients who fulfill research eligibility criteria. The bulk of the data in the RPDR is laboratory results, medication orders, radiology procedures, demographic data and text reports (radiology, pathology, cardiology, and discharge summaries). There is some coded clinical data, derived from problem lists and billing diagnoses, but much clinical data remains archived in text files.

1.4.1.2 Identifying Institutional Needs

Despite the current utility of the RPDR to identify research candidates, there is growing interest in accessing data contained within the narrative clinical text to hone cohort searches. It is recognized that procedural reports, which include echocardiogram reports, Pap smear results, and colonoscopy reports, contain valuable clinical information. This information is not always available in problem lists or billing diagnoses and has been quite difficult to acquire from ambulatory care notes and discharge summaries.

Data from these narratives will undoubtedly help to both broaden and increase the specificity of current queries into the system. Importantly, because of the sheer volume of data in the system, it

* COSTAR is a clinical record system and has been running in the out-patient primary care environment at the MGH since 1979. The COSTAR M database can be queried using the MQL language that can be used to search for clinical codes, laboratory values, and text patterns.

is imperative that high levels of precision are achieved. For example, in a typical query, if 50,000 patients matching criteria are returned, even if the system performs at a precision of 95%, then 5% or 2500 of these patients will be incorrectly identified. While this is not a large number compared with 50,000, these 2500 patients are 2500 more patients for whom expensive, labor-intensive chart review must be performed.

1.4.2 Using Available Tools and Software

As in many other institutions, natural language support to access this information is minimal and natural language processing systems cannot be easily implemented. The goal of this endeavor is to build a method of information extraction that can be generalized to other procedural reports, but is also relatively easy to implement using publicly available software and tools.

1.4.2.1 The Unified Medical Language System Metathesaurus and the MetaMap Program

The Unified Medical Language System (UMLS) Metathesaurus is the largest thesaurus in the biomedical domain containing concepts from more than 100 vocabularies and classifications. It provides a representation of biomedical knowledge consisting of concepts classified by semantic type and both hierarchical and non-hierarchical relationships among the concepts. This knowledge has proved useful for many applications including decision support systems, information extraction, information retrieval and data mining.

MetaMap is a program developed at the National Library of Medicine (NLM) to map biomedical text to concepts in the Metathesaurus. Both the Metathesaurus and the strategy used by the MetaMap program have been extensively described and will not be detailed here [28, 29]. An overview of the process used by MetaMap is briefly described in 2.2

1.5 Case Study: Information Extraction from Echocardiogram Reports

As discussed in the previous sections, information extraction methods using templates—either pre-defined or 'discovered'-- have been successfully applied to procedural reports. In the system described here, the UMLS and MetaMap are used to identify semantic classes of terms, which are subsequently used to generate patterns for individual sentences. Because of the ritualistic way in which procedural reports are dictated and clinical findings are narrated, these patterns occur repetitively across the vast majority of sentences in a report type. These class-based patterns are then used to identify templates containing the conditions and values of interest. This method should apply to many types of procedural reports. Because of the ubiquitous need for cardiac data, and because echocardiogram reports are probably the most highly structured of the procedural reports, this initial development effort and its evaluation is based on echocardiogram reports.

2 Methods: Development of Extraction System

2.1 Pre-processing

2.1.1 Corpus Overview

703 echocardiogram reports were analyzed for this study – 483 reports from Newton-Wellesley Hospital (NWH) and 220 reports from Faulkner Hospital (FH). Echocardiograms are dictated by staff cardiologists and are downloaded in HL7 format* to the RPDR on a daily basis. These reports were chosen from 19 randomly selected days between June and December 2003. 295 reports from NWH were used to train the system. The remaining 188 reports from NWH and all 220 reports from FH were used to complete a test set of 408 reports.

2.1.2 Concept Selection

The following ten concepts were used as benchmark conditions to populate the database: mitral valve insufficiency, aortic valve stenosis, pulmonary hypertension, mitral valve prolapse, valvular vegetations (including endocarditis), cardiac shunt (including patent foramen ovale, left to right shunt and right to left shunt), intracardiac thrombus, ejection fraction, pericardial effusion (including cardiac tamponade) and left ventricular hypertrophy. These ten concepts were felt to span a range of conditions and categories of concepts that could be important to extract from both research and clinical perspectives. Except for the following, each of these terms is classified as a [DISEASE OR SYNDROME] in the UMLS: patent foramen ovale ([CONGENITAL ABNORMALITY]), Ejection Fraction ([DIAGNOSTIC PROCEDURE]), cardiac shunt ([PATHOLOGIC FUNCTION]), left to right/right to left shunt ([ACQUIRED ABNORMALITY]).

* HL7 is a messaging standard for the electronic interchange of clinical, financial and administrative information among independent health care oriented computer systems; *e.g.*, hospital information systems, clinical laboratory systems, and pharmacy systems.

2.1.3 HL7 Parsing

A publicly available perl HL7 toolkit that provides a lightweight perl API for manipulating, sending and receiving HL7 messages was used to extract the note element from the HL7 message [30]. A few subroutine additions were necessary to extract the actual note. Figure A is an example of a single HL7 message containing an echocardiogram report.

2.1.4 Section Parsing

Depending on the source hospital, echocardiogram reports contain a variable number of sections. Reports from NWH generally had three sections: a coded field section, a description section, and an impression section. Reports from FH had a single impression section. Coded field regions were excluded from further analysis. For those reports with more than one section, while some of the information was redundant across the sections, this was not always true. A given sentence could be repeated across the sections in the same way, in different ways, or not at all. Regular expressions were used to extract the description section (if one existed) and the impression section. The narrative section of the echocardiogram report in Figure A is shaded grey. In those reports with a description section, the impression section occurred immediately after the description section and was indicated by "IMPRESSION" or "CONCLUSION" or "SUMMARY".

```
MSH|^~\&|MT^CAR|XXX|Datagate||111||XXX^T04|2
222222|T|2.3.1|111□PID||111^^^XXXX|||Lincoln
^Abraham^L^^^^L^A|MEDT||||ER-  Chest  Pain-
TXA|1|CAR|FT|||111|111|200310070946|1|EKG.MO
RL^SMITH^JOHN^^^^^^MT^L^^^^XXX^A|5555555555|
||||PA|||||^^^^^^^^^^^^^^□OBX|1|FT|ECHO&ZCR^
ECHOCARDIOGRAM     REPORT^XXX-CARM-OERPT|1|A
Hospital                      NAME: Lincoln,
Abraham             \.br\1 Washington Street
PHY:  Jefferson,  Thomas  M.D.  EXAM  DATE:
06/09/03                      \.br\CARDIOLOGY
DEPARTMENT                     STATUS: DIS IN
\.br\ORDERING PHY: Jefferson,Thomas F. M.D.
and              \.br\                  \.br\
ECHOCARDIOGRAM REPORT\.br\  \.br\Reason  for
ECHO        Study?         Chest         Pain
\.br\LEFT                           VENTRICLE
Right   Ventricular   Dimension(9-26).    wnl
\.br\
Left    Atrial    Dimension(19-40)......   41
\.br\Wall           thickness            (mm)
Aorta,     Dimension(20-37)...........     39
\.br\ Septum(6-11).. 11         Posterior(6-
11)..  11        Aortic  Valve  Excursion(15-
26).....  nml      \.br\Wall  Excursion  (mm)
Mitral    Valve    Excursion(20-35).....  nml
\.br\ Septum(5-12)..            Posterior(9-
14)..              \.br\Chamber   Dimension
EJECTION FRACTION: 65    %\.br\ Diastole(37-
56)...... 48         Systole..  30      \.br\
\.br\Minor  Axis  Shortening:          %\.br\
\.br\45  year  old  gentleman  with  positive
family        history        of      coronary
artery\.br\disease  and  atypical  chest pain
to   evaluate   for   any   evidence  of  wall
motion\.br\abnormalities.\.br\  A  trace  of
aortic  insufficiency.   There is no evidence
of   mitral   valve   prolapse.    No  mitral
insufficiency    is\.br\noted.    The    left
ventricle is normal in diastolic dimensions,
wall  thickness  and systolic\.br\function. No
wall  motion  abnormalities  are  noted.  The
ejection        fraction       is   estimated
at\.br\65%.\.br\   The   right   ventricular
systolic    pressure    is    calculated   at
34\.br\mmHg,    assuming    a   right  atrial
pressure of 10 mmHg. There is no evidence of
pericardial\.br\effusion. An echo free space
is  noted  around  the  heart,  consistent with
pericardial  fat.\.br\  \.br\IMPRESSION:\.br\
\.br\        1.      Normal  left  ventricular
dimensions,   wall   thickness   and  systolic
function.  The\.br\                  ejection
fraction is estimated at 65%.\.br\        2.
Mildly  dilated  aortic  root.  Trace  aortic
insufficiency is noted.\.br\    3.    Normal
right    ventricular    function   and  normal
pulmonary pressures.\.br\ \.br\ \.br\ \.br\
OEORD||11111|D□PR1|2||DE^DOPPLER         ECHO
(2011)^XX-CUM-
OEORD||11111|D□PR1|3||CFI^COLOR FLOW IMAGING
(2013)^XX-CUM-OEORD||11111|D
```

Figure A: Sample HL7 message containing an echocardiogram report from NWH. The area in black print is the echocardiogram report. A coded area, description section and impression section are present in the report.

2.2 Concept Mapping [28, 29] and Sentence Reconstruction

Developed at the National Library of Medicine (NLM), the MetaMap Transfer (MMTx) program is a free resource available by license. The program maps narrative text to concepts in the UMLS Metathesaurus. Briefly, MMTx passes a document into a module that *tokenizes* the text into sections containing sentences. Sentences are the tokenized into *word tokens*. The sentences can be passed to a part of speech tagger client (if one is implemented) to assign part-of-speech tags to each word.[*] If no part-of-speech tagger is implemented, each word token is compared with terms from the SPECIALIST LEXICON and the part-of-speech is retrieved. Up to three words can be concatenated to generate the best match. For example, the following sentence:

```
"There is no evidence of mitral valve."
```

Would be tokenized into:

```
There
Is
No
Mitral
Valve
Prolapse
.
```

After the lexical lookup process, this would become:

```
There
Is
No
Mitral Valve Prolapse
.
```

Each *term* is a *lexical element* made up of word tokens. Each lexical element is associated with its part-of-speech tag. Using these tags, noun phrases are identified and mapped to concepts in the UMLS. Variants, including synonyms, spelling derivations, acronyms and abbreviation are retrieved in a variant generation module. Each variant is marked with a cost or distance of how many transformations it took to get from the original form to the variant form. Phrases and their variants are identified and mapped to UMLS concepts where applicable.

A Java API provides a way to manipulate the MMTx process and its output to serve individual needs. Because of the inconsistent format of echocardiogram reports generated at NWH, document sections were parsed separately and then input by section to the MMTx processor. MMTx output was manipulated so that only the best mapping for each noun phrase as well as lexical information for each term is output.

[*] MMTx does not have a part of speech tagger, but does include hooks to a tagger via a programming interface. In our case, in order to use only what is already available, no part-of-speech tagger was used.

Each sentence was then reconstructed so that the position of the lexical element in the sentence served as a key to the details of that term. Each lexical element position ("4") was assigned its associated term ("mitral valve prolapse"), the part-of-speech ("noun"), the mapped concept ("mitral valve prolapse") and the semantic type of that concept ([DISEASE OR SYNDROME]). Except in the case of [VALUE], no concept or semantic type was assigned if MMTx did not map a term. [VALUE] assignments are discussed in the following section. Figure B is an example of how our sample sentence is reconstructed.

Key	Term	POS	Semantic Type
1	THERE	adv	null
2	IS	aux	null
3	NO	value	[VALUE]
4	MITRAL VALVE PROLAPSE	noun	[DISEASE OR SYNDROME]
5	.	punctuation	null

Figure B: Reconstruction of sentences involved mapping concept and lexical information to all lexical elements of the sentence.

2.3 Post-processing

A program in the Perl computer language was written to accomplish the following three tasks: identification of values, template recovery, and template matching and condition-value extraction.

2.3.1 Identification of Values

There is a general value scheme used by cardiologists to attribute degrees of findings diagnosable by echocardiograms (e.g. "mild | moderate | severe mitral valve insufficiency"). However, it was felt that the myriad variations (e.g. trace vs. mild vs. minimal vs. insignificant) are sufficiently subjective that an arbitrary mapping scheme may not be generally applicable. Extracting the values literally would allow each user to specify search values in their inclusion criteria.

Fortunately, while the semantics behind a particular value term is not always clear, the actual terms used to assign value comes from a fairly limited domain. While some values used in echocardiogram reports mapped to a QUANTITATIVE concept in the UMLS (e.g. "moderate"), some terms (e.g. "trace", "no") did not. Because of this variation, values were mapped in the post-processing phase using a separate VALUE LEXICON. The following terms were identified as a [VALUE]: *trace*, *mild*, *moderate*, *severe*, *insignificant*, *trivial*, *small*, *large*, *minimal*, *marked*, *slight*, *borderline*, *significant*, *modest*, *critical*, *substantial*, *less*, *very*, *neither*, *without*, *no*, *not* and *absent*. The adverb form of each [VALUE] (e.g. "moderately") was also permitted. *Value phrases* were identified when [VALUE] terms occurred in tandem (e.g. "moderately severe"), or when a value range was described (e.g. "trace to mild"). Values identified singly or as a value phrase were identified as a single [VALUE] term.

2.3.2 Template Recovery

The non-null *semantic types* of the lexical elements comprising a sentence were identified. These semantic types formed the *concept pattern* for that sentence. For example, the concept pattern for:

"There is [no] [mitral valve prolapse]."

is

[VALUE] [DISEASE OR SYNDROME]

Further pattern matching could then proceed in terms of these, more general semantic labels, rather than the surface word forms. The pattern above would also match:

"[Trace] [mitral valve insufficiency]."
"[No] evidence of [pericardial effusion]"
"There is evidence of [moderate] [aortic stenosis]."
...And so on.

Only those lexical elements with an associated semantic type contributed to the concept pattern. The concept patterns generated from all sentences in the training set were tallied and reviewed. Those patterns representing 3 or more unique sentences and containing at least one of the ten conditions were added to the system as a *template*, along with tags indicating the location of the [DISEASE OR SYNDROME] and the [VALUE].

In this review, it became apparent that even though a sentence did not match a complete pattern, part of the sentence would match certain *elemental patterns* with a high degree of specificity. Phrases, such as "consistent with [DISEASE OR SYNDROME]", appeared often and indicated the presence of that concept in the patient. It was clear that if sentences were mapped to these elemental phrases, the sensitivity of the systems would improve without appreciable reduction in precision. If we relied solely on the templates matching complete sentences, none of the remaining sentences would match and this data would be lost.

2.3.3 Template matching and concept-value pair extraction

Once all templates were added to the system, a concept pattern was generated for each new sentence. If no concept pattern could be identified, no information could be extracted. If a concept pattern did exist, then it was matched against the possible templates identified in the template recovery phase. At this point there were three possibilities-- the pattern of a given sentence could:

1. Match completely with one of the identified patterns,
2. Match partially with one of the elemental patterns, or
3. Not match at all with any of the patterns.

In the first and second cases, condition-value pairs were extracted according to what the template dictated. In the third case, if no match existed, the information contained within that sentence was ignored.

Extracted condition-value pairs along with the document source, and section source were output to a Microsoft Access database for further analysis.

3 Methods: Evaluation

The evaluation of this application took place in two phases. In both phases, recall and precision were measured. In the first phase, the system's ability to extract data accurately was evaluated by comparing findings extracted by the system with findings extracted manually by the author. In the second phase, the ability to use this information to retrieve relevant documents was evaluated. While this system was not built to be an information retrieval system, this second evaluation was felt to be appropriate for two reasons:

1. The output of this system will eventually be used to identify patient cohorts—part of this process involves identifying documents meeting specified criteria
2. Oftentimes, concepts and values stated one way may not generate a pattern that matches a template. However, this concept and value may be repeated elsewhere in the same document in a different way. This different way may generate a pattern that does match a template and therefore does extract the appropriate concept and value which can be identified in the retrieval process.

3.1 Generation of the Gold Standard

The author is a board-certified internist and is familiar with the information contained in echocardiogram reports. Each of the ten concepts along with their values was extracted and input into a database table. The author, as the developer of this system, was not blinded to the way in which information would be extracted from the reports. However, for the ten test conditions, identifying the concept and their values was a straightforward process. Identifying "no pericardial effusion" is a much easier task than interpreting "coronary artery disease" from "segmental hypokinesis in the inferobasilar region". Given the nature of echocardiogram reports, which is to explicitly state the presence, absence, and degree of disease, in most cases, there was little ambiguity. In cases where ambiguity was inherent in the sentence, the condition took on the value that was found associated with it. For example, in the first example in Figure C, "pericardial effusion" was given a value of "small". In future implementations, this may need to be reconsidered or possibly, a certainty term will be added.

```
"There is a small echo free space anterior to the heart that is
consistent with either a small pericardial effusion or an epicardial
fat pad."
```

Figure C: Example of an ambiguous sentence.

3.2 Scoring Method

For each concept-value pair, a simple score was computed from two numbers. The *coverage* is a measure of how much a given concept pattern accounts for the observed pattern of a sentence:

$$\text{coverage} = \frac{\#\text{ of observed concepts}}{\#\text{ of template concepts}} x100$$

The *distance* gives an estimate of how closely associated a concept and value may be and uses lexical distance as a surrogate. A distance score of 1 indicates that the concept and its value are next to each other. Conversely, if the distance score is high, this means that the concept and value are far away from each other. This number cannot exceed the total number of lexical elements in a sentence.

The final score is computed as follows:

$$\text{score} = \frac{\text{coverage - distance}}{100}$$

This scoring method accounts for the different levels of coverage that a sentence can have, but at each level of coverage, specifies that the closer the concept and value are to each other, the more likely they are to be associated. . Then for each level of coverage, the further apart the concept and value is, the lower the score. The upper limit of this score is 0.99 (lower limit is 0) because the distance score must be at least one. The reason for this was in some ways practical. Given the nature of this procedure, we did not want to give anyone the impression that we are 100% sure of any concept-value pair.

Table 1: Mapping rules for test conditions.

Concept	Body Part	Finding
Mitral Valve Insufficiency	-	Mitral valve Insufficiency
	-	MR
	Mitral valve	Mitral valve Insufficiency Insufficiency, Unspecified Regurgitation, Unspecified MR
	Mitral	Mitral valve Insufficiency Insufficiency, Unspecified Regurgitation, Unspecified MR
Aortic Valve Stenosis	-	Aortic Valve Stenosis
	Aortic valve	Aortic Valve Stenosis Stenosis, Unspecified
	Aortic	Aortic Valve Stenosis Stenosis, Unspecified
Pericardial Effusion	-	Pericardial Effusion
	-	Cardiac Tamponade
	-	Tamponade
Ejection Fraction	-	Ejection Fraction
	-	EF
	-	Left Ventricular Ejection Fraction
Mitral Valve Prolapse	-	Mitral Valve Prolapse
	Mitral valve	Mitral Valve Prolapse Prolapse
	Mitral	Mitral Valve Prolapse Prolapse
Intracardiac Thrombus	-	Intracardiac Thrombus
	-	Mural thrombus
	-	Thrombus
Pulmonary Hypertension	-	Pulmonary Hypertension
Cardiac Shunt (including patent foramen ovale)	-	Right to Left Shunt
	-	Left to Right Shunt
	-	Cardiac Shunt
	-	Shunt
Patent Foramen Ovale, alone	-	Formen Ovale, Patent
	Atrial septum	Formen Ovale, Patent
Valvular Vegetations	-	Valvular Vegetations
	-	Endocarditis
	-	Vegetation
	-	Vegetations
	All valvular parts*	Valvular Vegetations Endocarditis Vegetation Vegetations
Left Ventriclular Hypertrophy	-	Hypertrophy, Left Ventriclular
	Left ventricle	Ventriclular Hypertrophy Hypertrophy

*All valvular parts: mitral valve, mitral, aortic valve, aortic, tricuspid valve, tricuspid, pulmonary valve, pulmonic

3.3 Mapping Rules

In order to evaluate the information extraction and information retrieval aspects of this system, a set of mapping rules was created in order to normalize the representation of the ten test conditions in the database. The [DISEASE] along with a [BODY PART] or [ANATOMICAL STRUCTURE] characterizes each condition. For example, the condition "mitral valve insufficiency" can be mapped from:

- "mitral valve insufficiency [DISEASE]",
- "mitral valve [BODY PART]" and "insufficiency [DISEASE]",
- "mitral [ANATOMICAL STRUCTURE]" and "insufficiency [DISEASE]".

Table 1 shows the mapping schema for the 10 test concepts.

The set of mapping rules for each concept was used to group entries by each parent concept. Ultimately, the goal is to use the concept hierarchy built into the UMLS to generate these mapping rules.

3.4 Information Extraction

Condition-value pairs related to the ten test conditions were identified from both the experimental set and from the gold standard using these mapping rules.

The following definitions were used:

1. *True positive*: Condition-value pair with matching document id, section type, concept and value in both sets.
2. *False positive*: Condition-value pair found only in the experimental set.
3. *False negative*: Condition-value pair found only in the gold standard.

Since the reference standard includes all discoverable condition-value pairs and by definition, does not include pairs that do not exist, *true negatives* cannot be identified.

Using these three parameters, recall and precision measures were calculated as follows:

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

3.5 Document Retrieval

Retrieval effectiveness is also measured using recall and precision; however, *documents* instead of *findings* are the unit of interest.

A system's ability to retrieve relevant documents is assessed with a recall measure:

$$\text{Recall} = \text{Sensitivity} = \frac{\text{\# of relevant documents returned}}{\text{total \# of relevant documents in the collection}}$$

Of course, a system can achieve a recall of 100% if all documents are returned. This measure is balanced with an assessment of accuracy. A system's accuracy is based on how many of the documents returned for a given query are actually relevant, which can be assessed by a precision metric:

$$\text{Precision} = \text{PPV} = \frac{\text{\# of relevant documents returned}}{\text{\# of documents returned}}$$

In addition to these measures, specificity was calculated in order to obtain the area under the receiver operating characteristic curve (AUC). The AUC provides a measure of how well the scoring method discriminates between cases and non-cases.

In this evaluation, a document was judged to be relevant if a condition's value matched one of those indicated in Table 2. For most conditions, if the disease was present in more than a 'mild' way, the document was retrieved. This was done in order to replicate a true-to-life situation in which queries to the RPDR will ask to find patients with a disease of a particular severity. For example, in order to find patients with aortic valve stenosis, those documents with aortic valve stenosis findings associated with values *moderate*, *moderate-severe*, *moderately severe* and *severe* were retrieved. The exception was for the condition 'valvular vegetations'. Since none of the test documents had a positive finding for this condition, all documents with 'no' evidence of vegetations was retrieved instead.

Table 2: Quantitative and qualitative values associated with test conditions.

Concept	Quantitative Values	Qualitative Values
Mitral Valve Insufficiency	1-2+	mild to moderate
	2+	moderate
	2-3+	moderate (2+/4+)
	3+	moderate to moderately severe
	4+	moderate to severe
		moderately severe
		moderately severe (3+ out of 4+)
		moderately severe to severe
		moderately severe tosevere (3-4+/4+)
		significant
		severe
		present
Aortic Valve Stenosis	1-2+	moderate
		moderately severe
		significant
		severe
		present
Pericardial Effusion		moderate
		significant
		present
Mitral Valve Prolapse		borderline
		mild
		minimal
		slight
		very slight
		present
Intracardiac Thrombus		present
Pulmonary Hypertension	44 mmHg	mild to moderate
	45 mmHg	moderate
	60 mmHg	moderate to severe
	65 mmHg	moderately severe
	74 mmHg	marked
	75 mmHg	severe
		significant
		present
Cardiac Shunt		minor
		very small
		small
		tiny
		present
Patent Foramen Ovale		minor
		very small
		small
		tiny
		present
Valvular Vegetations		no
		without
Left Ventricular Hypertrophy		mild to moderate
		mildly to moderately
		moderate
		marked
		severe
		present

3.6 Tools

Recall, precision and specificity measures were calculated using SAS version 8.0. Area under the ROC curve (AUC) was calculated using a perl script available under the GNU license.

4 Results

4.1 Corpus Description

Table 3 provides a description of the training and the test sets. Additional analysis was done on test sets from each hospital. On average the training set had 7-8 unique sentences per report; the entire test set had an average of 4-5. This disparity is largely because of the differences between the NWH corpus and the FH corpus. The NWH test corpus also had 7-8 unique sentences per report, whereas reports from FH had 2-3. In fact, although FH had a greater number of documents in the test set, the number of unique sentences, unique words and generated patterns were significantly less than the corresponding numbers in the NWH test set. As discussed in the Introduction, procedural reports invariably contain some degree of computer-generated text. While this is certainly known to be true of echocardiogram reports, this analysis demonstrates that the extent of this practice is likely institutional, and may even be author-specific. Importantly, the number of unique sentences occurring singly and the mismatch between patterns and unique sentences suggests that there is clearly opportunity for authors to revise and that this opportunity is often taken.

Table 3: Description of corpus.

	Train		**Test**					
Corpus Characteristic	**NWH (n=295)**		**NWH+FH (n=408)**					
			NWH (n=188)		**FH (n=220)**		**NWH+FH**	
	Total	**Single**	**Total**	**Single**	**Total**	**Single**	**Total**	**Single**
Unique sentences	2296	1720	1474	1068	494	242	1939	1293
Unique words	1390	518	1108	339	576	153	1350	393
Patterns	1255	885	863	581	321	146	1130	698
Unmatched sentences	-	105	-	74	-	35	-	109

4.2 Patterns

Table 4 shows the number of total patterns isolated and the number of templates identified in the template recovery phase. Each pattern was categorized according to the number of sentences it can map and was grouped as indicated in the first column. The number of sentences mapped by the templates in that category is shown in the last column. For example, only 10 patterns matched ≥50 sentences each. Of these 10 patterns, only 4 mapped to sentences containing information on at least one of the 10 test conditions. Those 4 patterns matched 903 sentences. This demonstrates the power of this technique to extract large number of sentences with comparatively fewer patterns. A total of 55 templates were used to extract condition-value pairs from 1390 sentences.

Table 4: Pattern analysis

Sentence Matches/Pattern*	# patterns total	# patterns used	# sentences
≥ 50	10	4	903
10-49	62	14	317
5-9	59	13	89
3-4	91	24	81
Total	160	55	1390

* Number of sentences matching the number of concept patterns indicated in the training corpus.

4.3 Information Extraction

Using the largest test set (all 55 patterns used), 1258 condition-value pairs were extracted. Table 5 shows the distribution of extracted pairs over the test conditions. The final column is a measure used simply to show how the effectiveness of this method varies from condition to condition. This method works best for extracting "Ejection Fraction" and worst for non-patent foramen ovale "cardiac shunts." This provides a measure of how sentence structure varies from disease to disease.

Table 5: Comparison of extraction efficiency among test conditions.

Condition	n		FP	FP*	(O-FP*)/E
	Expected (E)	Observed (O)			
Mitral Valve Insufficiency	389	326	2	0	0.84
Aortic Valve Stenosis	115	89	0	0	0.77
Ejection Fraction	450	415	10	3	0.92
Mitral Valve Prolapse	50	36	2	0	0.72
Pulmonary Hypertension	45	31	0	0	0.69
Left Ventricular Hypertrophy	139	18	0	0	0.13
Pericardial Effusion	330	323	0	0	0.98
Cardiac shunt (excluding patent foramen ovale)	36	2	0	0	0.06
Patent Foramen Ovale, alone	19	8	0	0	0.42
Valvular Vegetations	14	4	0	0	0.29
Intracardiac Thrombus	9	6	0	0	0.67
	1596	1258	14	3	

*False positives included only semantic false positives

The following four tables show the results of the information extraction component of this evaluation. Analysis was performed on sequentially larger template bases in order to demonstrate the utility of using even small numbers of templates to reliably extract pertinent information. All tables report recall and precision, but with slightly adjusted definitions. Table 6 and Table 7 show the recall and precision over all extracted concepts. Table 7, however, uses an adjusted definition for false-positives. As discussed in the Methods, literal value assignments were used. Because of the way that value assignments were discovered, however, atypical value presentations, e.g. "(mild) 1+/4+", were extracted slightly anomalously. The system would extract "mild) 1+/4+" in the previous example. In the database system, however, a rudimentary string matching procedure matched values together. Therefore, unless a direct string match was successful, the concept-value pair was deemed to be false-positive. While some of the pairs identified in the overall analysis were true false-positives (as might occur when the true value is "no significant" but the system extracts "significant"), others were semantically equivalent, but lexically unequal (e.g. "(mild) 1+/4+" v. "mild) 1+/4+"). Table 7 shows the overall results in which false-positives are the 'true' semantic false-positives. By the most conservative estimate—using the values in Table 6-- the system using all patterns was able to recall 78% of the relevant findings (95% CI, 76% to 80%), with a precision of 99% (95% CI, 98%-99%).

Table 6: Overall recall and precision (including lexical false positives).

Pattern Category	Train (%, [95% CI])		Test (%, [95% CI])					
	NWH		NWH		FH		NWH + FH	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
≥ 50	69 [66, 71]	99 [98, 100]	65 [62, 68]	99 [99, 100]	51 [48, 55]	99 [97, 100]	59 [57, 61]	99 [99, 100]
≥ 10	82 [80. 84]	99 [99, 100]	78 [76, 81]	99 [99, 100]	71 [67, 74]	98 [97, 99]	75 [73, 77]	99 [98, 99]
≥ 5	85 [83, 87]	99 [99, 100]	81 [78, 83]	99 [99, 100]	72 [68, 75]	98 [97, 99]	77 [75, 79]	99 [98, 99]
≥ 3	87 [85, 88]	99 [99, 100]	83 [80, 85]	99 [99, 100]	72 [69, 76]	98 [97, 99]	78 [76, 80]	99 [98, 99]

Table 7: Overall recall and precision with true false positives.

Pattern Category	Train (%, [95% CI])		Test (%, [95% CI])					
	NWH		NWH		FH		NWH + FH	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
≥ 50	69 [67, 72]	99 [99, 100]	65 [62, 69]	100 [100, 100]	52 [48, 56]	100 [100, 100]	59 [57, 62]	100 [100, 100]
≥ 10	83 [81, 85]	99 [99, 100]	79 [76, 81]	100 [100, 100]	72 [68, 75]	100 [99, 100]	76 [74, 78]	100 [99, 100]
≥ 5	85 [83, 87]	100 [99, 100]	81 [79, 84]	100 [100, 100]	73 [70, 76]	100 [99, 100]	77 [75, 80]	100 [99, 100]
≥ 3	87 [85, 89]	100 [99, 100]	83 [80, 85]	100 [100, 100]	73 [70, 77]	100 [99, 100]	79 [77, 81]	100 [99, 100]

In order to see how the system would perform if only the top-scoring condition-value pairs were retrieved, the following two tables shows recall and precision measures when only condition-value pairs with a score of 0.99 are used. Table 8 shows recall and precision measures using the full false-positive data and Table 9 again shows the analysis with the lexically mismatched false-positives re-designated as true-positives.

Table 8: Recall and precision for entries with score of 0.99.

Pattern Category	Train (%, [95% CI])		Test (%, [95% CI])					
	NWH		NWH		FH		NWH + FH	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
≥ 50	23 [21, 25]	99 [99, 100]	23 [21, 26]	100 [99, 100]	28 [24, 31]	100 [100, 100]	25 [23, 27]	100 [99, 100]
≥ 10	30 [27, 32]	100 [99, 100]	29 [26, 32]	100 [99, 100]	28 [25, 32]	100 [100, 100]	29 [27, 31]	100 [99, 100]
≥ 5	33 [31, 35]	100 [99, 100]	34 [31, 37]	100 [99, 100]	29 [26, 33]	100 [100, 100]	32 [30, 34]	100 [99, 100]
≥ 3	36 [33, 38]	100 [99, 100]	36 [33, 39]	100 [99, 100]	29 [26, 33]	100 [100, 100]	33 [31, 35]	100 [99, 100]

Table 9: Recall and precision for entries with score of 0.99 and with true false positives.

Pattern Category	Train (%, [95% CI])		Test (%, [95% CI])					
	NWH		NWH		FH		NWH + FH	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
≥ 50	23 [21, 25]	100 [100, 100]	23 [21, 26]	100 [100, 100]	28 [24, 31]	100 [100, 100]	25 [23, 28]	100 [100, 100]
≥ 10	30 [28, 32]	100 [100, 100]	29 [26, 32]	100 [100, 100]	28 [25, 32]	100 [100, 100]	29 [27, 31]	100 [100, 100]
≥ 5	33 [31, 36]	100 [100, 100]	34 [31, 37]	100 [100, 100]	29 [26, 33]	100 [100, 100]	32 [30, 34]	100 [100, 100]
≥ 3	36 [33, 38]	100 [100, 100]	36 [33, 39]	100 [100, 100]	29 [26, 33]	100 [100, 100]	33 [31, 36]	100 [100, 100]

4.4 Document Retrieval

Table 10 shows the results of the document retrieval phase of this analysis. For each of the 10 conditions listed, recall, precision, sensitivity and AUC are calculated for each pattern category. In cases where the result was the same across all pattern categories, a single result is listed. Notably, in the test set, no cases of 'intracardiac thrombus' were identified across all pattern categories, and no cases of 'left ventricular hypertrophy' were discovered when using patterns that matched only 50 or more sentences. Because data acquired from the system will ultimately be used in document and patient retrieval, this preliminary analysis was done to evaluate how this additional data might contribute to patient retrieval. Median recall across the ten conditions was 36% (range, 0% to 93%). The system retrieved no documents for two of the ten conditions; median precision for the remaining eight conditions was 100% (range, 92% to 100%).

Table 10: Document Retrieval by Clinical Condition

Condition	Pattern Category	Prevalence	Precision	Recall	Specificity	AUC	p
Mitral Valve Insufficiency	_ 50	0.15	92 [85, 100]	77 [67, 88]	99 [98, 100]	0 88 [0.82, 0.94]	<1 0
	_ 10			77 [67, 88]		0 88 [0.82, 0.94]	
	_ 5			79 [69, 89]		0 89 [0 83, 0 95]	
	_ 3			79 [69, 89]		0 89 [0.83, 0.95]	
Aortic Valve Stenosis	_ 50	0 03	100 [100, 100]	67 [36, 97]	100 [100, 100]	0 83 [0.67, 1 00]	<1.0
	_ 10						
	_ 5						
	_ 3						
Pericardal Effusion	_ 50	0.04	100 [100, 100]	93 [81, 100]	100 [100, 100]	0.97 [0 90, 1.00]	<1.0
	_ 10						
	_ 5						
	_ 3						
Mitral Valve Prolapse	_ 50	0.05	100 [100, 100]	55 [33, 77]	100 [100, 100]	0 78 [0 62, 0 93]	<1.0
	_ 10			55 [33, 37]		0.78 [0 62, 0 93]	
	_ 5			65 [44, 86]		0 82 [0 71, 0 94]	
	_ 3			75 [56, 94]		0 88 [0 77, 0.98]	
Intracardiac Thrombus	_ 50	0.002	-	0	100 [100, 100]	-	-
	_ 10		-			-	-
	_ 5		-			-	-
	_ 3		-			-	-
Pulmonary Hypertension	_ 50	0 07	100 [100, 100]	25 [9, 41]	100 [100, 100]	0.63 [0.53, 0.72]	<1.0
	_ 10			46 [28, 65]		0.73 [0.63, 0 84]	
	_ 5			46 [28, 65]		0 73 [0.63, 0 84]	
	_ 3			46 [28, 65]		0.73 [0 63, 0.84]	
Cardiac Shunt	_ 50	0 03	100 [100, 100]	31 [6, 56]	100 [100, 100]	0 65 [0.52, 0 79]	<1 0
	_ 10						
	_ 5						
	_ 3						
Patent Foramen Ovale	_ 50	0.03	100 [100, 100]	36 [8, 65]	100 [100, 100]	0 68 [0 53, 0.84]	<1.0
	_ 10						
	_ 5						
	_ 3						
Valvular Vegetations	_ 50	0.03	100 [100, 100]	9 [8,26]	100 [100, 100]	0 55 [0 44, 0.65]	<1.0
	_ 10			9 [8, 26]		0 55 [0 44, 0 65]	
	_ 5			27 [1, 54]		0.64 [0.49, 0 78]	
	_ 3			27 [1.54]		0.64 [0 49, 0.78]	
Left Ventricular Hypertrophy	_ 50	0.09	-	0	100 [100, 100]	-	-
	_ 10		100 [100, 100]	13 [2, 24]	100 [100, 100]	0 57 [0 49, 0 64]	<1.0
	_ 5						
	_ 3						

5 Discussion

5.1 Concept-Pattern Templates as a Solution

There were two goals for this project. First, a simple implementation was sought because of limited local expertise and support. This was accomplished using the publicly available UMLS Metathesaurus vocabulary and the MetaMap concept-mapping system. Previous work has shown that there are sentence types and word class patterns that exist with high frequency and which can be used to extract findings and their values with high precision. This evaluation has shown that templates based on semantic patters can lead to productive extraction of clinical data from echocardiogram reports. Second, for reasons discussed in the Background, we sought to develop a precise system, rather than one that would capture more information. In terms of information extraction, perfect or near perfect precision was achieved at reasonable levels of recall. Importantly, this high level of precision was maintained even when fewer patterns were used. This suggests that with one can potentially extract a significant amount of information with minimal review of the most frequently occurring patterns.

It is understood, however, that the tolerance for both recall and precision can vary according to need and to the conditions of interest. For example, we can see that "intracardiac thrombus" is relatively rare with prevalence in the test sample of 0.2%. If a researcher were to query the system and request a high level of precision, this query may not yield many cases. In fact, it may be that any mention of this disease may be sufficient evidence of its positive status in a given patient. In cases like these, researchers may elect to ratchet down the level of precision required to simply include as many patients as possible. The simple scoring method proposed here demonstrates reasonably good ability to discriminate between documents with positive findings and those with absent or negative findings depending on the disease. A score of 0.99 achieves the highest level of precision and consequently the lowest recall. As the score threshold decreases, recall increases substantially, but precision predictably falls. The scores allow users of the system to specify the level of precision and recall they desire for a particular request.

While the system's performance on information extraction tasks was actually quite good, its performance in document retrieval suffered a much wider range of behavior. The bulk of this discrepancy is likely due to the conditions studied in each analysis. In the information extraction analysis, all condition-value pairs are looked at—the majority of which are related to ejection fraction and to the valvular disorders (mitral valve insufficiency and aortic valve stenosis), and the bulk of which are negative findings, which the information retrieval system was not asked to identify. Therefore, although the ability of this method to extract information is very good, information useful for the criteria used in the document retrieval process was not uncovered for these less common conditions. Note, however, that for those diseases that are more common, e.g. mitral valve insufficiency, aortic valve stenosis, and pericardial effusion, both the system and the scoring method performed as well in document retrieval as they did in the information extraction analysis.

5.2 Limitations

While this method shows promise, there are certain limitations of this system and of its analyses that should be addressed: Some of the limitations of this system have been found in other information extraction systems, which have also been able to achieve relatively good precision, but poorer recall. Because of the shallow parsing, most information extraction systems can only extract what is explicit. To get the rest of the information requires inference. Handling the most common phenomena gets you to 60% relatively quickly—getting to 100% requires handling increasingly rare phenomena [11]

5.2.1 Gold Standard

The author generated the gold standard. For most cases, this is not an issue since findings and their associated values are usually explicitly stated in procedural report. In general, however, it is better to allow someone (or more) not involved in the development process to generate the gold standard. Hripscak et al. performed a reliability study for evaluating information extraction from radiology reports to assess the reliability of a reference standard for an information extraction task. 24 physician raters from two sites and two specialties judged whether clinical conditions were present based on reading chest radiograph reports. They found that one to two raters were needed to achieve a reliability of 0.70, and six raters, on average were required to achieve a reliability of 0.95. They concluded that once the reliability of a specific rater is confirmed, it would be possible for that rater to create a reference standard reliable enough to assess aggregate measures on a system. Six raters would be needed to create a reference standard sufficient to assess a system on a case-by-case basis [31].

5.2.2 Disease Selection

Many common and a few uncommon reported conditions were chosen to evaluate this system. Uncommon positive conditions may be prone to greater narrative and descriptive license than common, negative findings. One can surmise that there is less opportunity for canonical descriptions to develop as a condition becomes less common or as the finding becomes 'more positive'. Chances of extracting these findings may increase if we look at patterns matching less than three sentences, or if more elemental patterns are discovered and used. Additionally, if we examine a larger number of echocardiogram reports from which templates are identified, it may be that uncommon conditions do have a useful pattern, which can only be revealed by broadening the scope of inquiry to include many more reports.

5.2.3 False Positives

False positives in the information extraction stage occurred when a concept was correctly identified as existing, but the value assigned to that concept was incorrect. Related these false-positives are those that arose in the document retrieval phase. Because the value associated with a concept was incorrectly identified, retrieval of documents containing these concepts and values were also falsely identified. The false positives identified in both evaluations result from the same processes and can be clustered into three categories:

The simplest, and most easily correctable, are those in which the string matching process failed to match two essentially equivalent strings. This can be resolved by modifying either the way that values are identified in the value identification process or it can be dealt with at the database level with some approximate string matching technique.

The second category of false positives resulted from intervening words. The statement "insignificant and probably small pericardial effusion" yields a value of "insignificant small" in the system, while the true value is more exactly "insignificant and probably small". Again, from the reader's perspective these two values may be the same (except that in the second case, the certainty on the second value assessment is less clear). This problem is somewhat more difficult to resolve in the present schema, as there is no way currently to assess certainty in a statement. The assumption up to this point has been that if there is a specific value, then the certainty must be high. Though this is true for many of the sentences, there are a sufficient number of sentences for which this cannot be assumed and could lead to misleading results. The addition of a certainty 'modifier' might be considered.

The third category of false positives is related to the second in that the intervening words between two values associated with a concept could potentially alter the value of a concept. Consider the following two sentences:

1. "[Aortic sclerosis] [without] [aortic stenosis] [or] [significant] [aortic regurgitation]."
2. "[Aortic sclerosis] with [significant] [aortic stenosis] [and] [no] evidence of [aortic regurgitation]."

Based on semantic patterns, these two sentences are equivalent:

[DISEASE] [VALUE] [DISEASE] [CONJ] [VALUE] [DISEASE]

However, in the first example, "aortic stenosis" has a value of "without" and "aortic regurgitation" has a value of "without significant." In the second example, "aortic stenosis" is "significant" and "aortic regurgitation" has a value of "no". Because both these sentences use the same sentence template, information from only one of these sentences will be extracted correctly. In cases such as these, template matching based on semantic type only is clearly insufficient. It is not enough to know that there is a conjunction in between these two concepts—it is necessary to know the kind of conjunction ("and" vs. "or") and to understand the way that values distribute over these concepts based on the type of conjunction present. As mentioned in Methods, lexical, and conceptual term information were collected to facilitate future system modifications.

5.3 Future Directions

Based on the evaluation of this prototype system, much work lies ahead for the future enhancement and further evaluation of this system.

5.3.1 Normalization of Values, Value Assessments and Assigning Certainty

The goal of assigning values to concepts is to assign values that the author intended. The best way to do this is to directly use the value string that the author used. As we can see from the evaluation process however, while using the exact string maintains the integrity of the data, this may not offer the best solution from the retrieval point of view. It may be that values will need to be normalized in some way using some mapping procedure or that templates containing more than one value concept need to be evaluated slightly differently than their simpler one-concept counterparts, possibly by adding a modifier.

5.3.2 Implementation of a Part-of-Speech Tagger

Part-of-speech tagging will allow improved identification of noun phrases and will allow improved sentence parsing. Syntactic information was not used in this system partly because it has not been shown to contribute extensively in information extraction efforts, but also because the lexical information provided by the LEXICON without the help of a tagger was deemed to be not reliable enough to use. It will be helpful to see how the addition of syntactic information influences performance.

5.3.3 Automated Pattern Discovery

Most of the work in this area has involved learning patterns from annotated documents—documents which have been marked up to indicate the relevant information, or documents for which extraction templates have been prepared by hand. Riloff described a learning procedure in which documents only had to be marked for relevance to the extraction task [32]. Yangarber et al. introduced an algorithm for learning good-quality patterns automatically from a large un-annotated corpus that begins with a few seed patterns strongly associated with the topic of interest [33]. These seed patterns are in turn used to identify relevant documents from which candidate patterns are proposed. Each candidate pattern is then matched against the entire corpus to check how closely it correlates statistically with the relevant document set: the pattern is considered "good" if it matches on the relevant documents substantially more often than on the non-relevant documents. The most strongly correlated patterns are then added to the seed patterns, and the cycle is continued. Automated pattern discovery may be a critical adjunct especially as the number of conditions of interest increase. The evaluation of this system has shown that manually identified patterns based on semantic type can be used with success to extract information from echocardiogram reports. The next step is to see if automated techniques can speedily recover appropriate patterns in order to facilitate this process.

5.3.4 Evaluation in the RPDR

The ultimate purpose of this information extraction system is to extract condition-value pairs that can be used to hone searches for patients fulfilling research criteria. While this evaluation suggests that this extracted data may be useful, evaluation in the RPDR setting where a larger number of records are scanned will be necessary. This level of evaluation will provide a better assessment of the system's ability to scale with acceptable recall and precision under production conditions.

5.3.5 Expansion to Other Procedural Reports

While echocardiogram reports generate much needed clinical information that are useful for researchers, there exist a wide variety of other medical narrative documents of similar constructs to which this method can be applied and studied.

5.4 Conclusions

Despite the enormous cache of data residing in clinical care systems, certain information is simply not accessible. Manual coding systems, though evolving, currently do not match the speed and simplicity of dictating narrative reports. Language technology, although promising, is still not sufficiently reliable to be widely applied in a variety of medical domains.[21] Exploiting narrative clinical information is not trivial.

We were interested in developing a practical free-text processing system that would be relatively easy to implement and amenable to application outside the test domain with few modifications. This effort was made in order to meet the demands of a large community of researchers requiring specific clinical information to further limit their domain of inquiry. Building an information extraction system based on semantic type has been shown to be effective in several domains.

This paper offers further evidence of the utility of concept-based templates in information extraction systems for clinical reporting and shows how publicly available tools can be leveraged to facilitate the development process.

As long as natural language processing applications are still new and relatively untested, storing extracted data in parallel with data collected by existing applications allow the information to be used immediately by those who wish to put faith in it, or not used by those who still have doubts. Until these strategies reach a level of reliability that is both acceptable for clinical care and is widely available, this is a solution for those who wish to have access to at least some of this information. Even at a recall of 50%, in a database of 1.4 million patients, this translates into 700,000 patients about whom we now know a little more about.

References

1. Hripsack, G., et al., *Unlocking Clinical Data from Narrative Reports.* Annals of Internal Medicine, 1995. **122**(9): p. 681-688.
2. Safran, C., et al., *Guidelines for management of HIV infection with computer-based patient's record.* Lancet, 1995. **346**(8971): p. 341-346.
3. Rector, A.L., W.A. Nowlan, and S. Kay, *Foundations for an electronic medical record.* Methods Inf. Med., 1991. **30**(3): p. 179-186.
4. Tange, H.J., et al., *Medical narratives in electronic medical records.* International Journal of Medical Informatics, 1997. **46**(1): p. 7-29.
5. McDonald, C.J., et al., *The Regenstrief Medical Record System: twenty years of experience in hospitals, clinics and neighborhood health centers.* MD Computing, 1992. **9**: p. 206-217.
6. McDonald, C.J., *The barriers to electronic medical record systems and how to overcome them.* Journal of the American Medical Informatics Association, 1997. **4**(3): p. 213-221.
7. Sager, N., et al., *Medical Language Processing: Computer Management of Narrative Data.* 1 ed. 1987: Addison-Wesley Publishing Company. 348.
8. Johnson, S.B. and C. Friedman. *Integrating Data from Natural Language Processing into a Clinical Information System.* in *Annual Symposium of the American Medical Informatics Association.* 1996. Washington, DC: Hanley & Belfus, Inc.
9. Lin, R., et al. *A free-text processing system to capture physical findings: Canonical phrase identification system (CAPIS).* in *Fifteenth Annual Symposium on Computer Applications in Medical Care.* 1991. Washington, DC: McGraw-Hill, Inc.
10. Jurafsky, D. and J.H. Martin, *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition.* 1 ed. 2000, Upper Saddle River, New Jersey: Prentice Hall. 934.
11. Hobbs, J.R., *Information extraction from biomedical text.* Journal of Biomedical Informatics, 2002. **35**(4): p. 260-264.
12. Appelt, D., et al. *FASTUS: A finite-state processor for information extraction from real-world text.* in *Proceedings of the Thirteenth International Joint Conference on Aritifical Intelligence.* 1993.
13. Grishman, R. and B. Sundheim. *Message understanding conference-- 6: A brief history.* in *Proceedings of the Sixteenth International Conference on Computational Linguistics.* 1996.
14. Grishman, R., S. Huttunen, and R. Yangarber, *Information extraction for enhanced access to disease outbreak reports.* Journal of Biomedical Informatics, 2002. **35**(4): p. 236-246.
15. Friedman, C., et al., *A general natural-language text processor for clinical radiology.* Journal of the American Medical Informatics Association, 1994. **1**(2): p. 161-174.
16. Krauthammer, M. and G. Hripsack. *A knowledge model for the interpretation and visualization of NLP-parsed discharge summaries.* in *Annual Symposium of the American Medical Information Association.* 2001. Washington, DC: Hanley & Belfus, Inc.
17. Xu, H. and C. Friedman. *Facilitating Research in Pathology using Natural Language Processing.* in *Annual Symposium of the American Medical Informatics Association.* 2003. Washington, DC: Hanley & Belfus, Inc.
18. Zingmond, D. and L. Lenert, *Monitoring free text data using medical language processing.* Comput. Biomed. Res., 1993. **26**: p. 467-481.
19. Canfield, K., et al. *Database capture of natural language echocardiographic reports: A Unified Medical Language System Approach.* in *Thirteenth Annual Symposium on Computer Applications in Medical Care.* 1989. Washington, DC: IEEE Computer Society Press.
20. Haug, P.J., D.L. Ranum, and P.R. Frederick, *Computerized extraction of coded findings from free-text radiologic reports.* Radiology, 1990. **174**(2): p. 543-548.
21. Mikkelsen, G. and J. Aasly, *Manual semantic tagging to improve access to information in narrative electronic medical records.* International Journal of Medical Informatics, 2002. **65**: p. 17-29.
22. Brown, P.J.B. and P. Sonksen, *Evaluation of the quality of information retrieval of clinical findings from a computerized patient databse using a semantic terminological model.* Journal of the American Medical Informatics Association, 2000. **7**: p. 392-403.

23. Barrows, R.C., M. Busuioc, and C. Friedman. *Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches.* in *Annual Symposium of the American Medical Informatics Association.* 2000. Los Angeles, CA: Hanley & Belfus, Inc.

24. Dick, R.S., *The computer based patient record.* 1997.

25. van Ginnecken, A.M., J. van der Lei, and P.W. Moorman. *Towards unambiguous representation of data.* in *Sixteenth Annual Symposium on Computer Applications in Medical Care.* 1992. Baltimore, MD: McGraw-Hill, Inc.

26. Murphy, S.N. and H.C. Chueh. *A security architecture for query tools used to access large biomedical databases.* in *Annual Symposium of the American Medical Informatics Association.* 2002. San Antonio, TX: Hanley & Belfus, Inc.

27. Murphy, S.N., et al. *Optimizing healthcare research data warehouse design through past COSTAR query analysis.* in *Annual Symposium of the American Medical Informatics Association.* 1999. Washington, DC: Hanley & Belfus, Inc.

28. Aronson, A.R. *Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program.* in *Annual Symposium of the American Medical Informatics Association.* 2001. Washington, DC: Hanley & Belfus, Inc.

29. Aronson, A.R., *MetaMap: Mapping Text to the UMLS Metathesaurus.* 1996.

30. http://hl7toolkit.sourceforge.net/.

31. Hripsack, G., et al., *A reliability study for evaluating information extraction from radiology reports.* Journal of the American Medical Informatics Association, 1999. **6**(2): p. 143-150.

32. Riloff, E. *Automatically generating extraction patterns from untagged text.* in *Proceedings of the Thirteenth National Conference on Artificial Intelligence.* 1996.

33. Yangarber, R., et al. *Automatic acquisition of domain knowledge for information extraction.* in *Proceedings of the Eighteenth Internation Conference on Computational Linguistics.* 2000.